

Encapsulated Models for Reasoning and Decision Support

Ivana Čače¹, John-Jules Ch. Meyer^{1,2}, C.R.C. Pieterman³, G.D. Valk³

¹ Alan Turing Institute Almere
Louis Armstrongweg 84
1311 RL Almere
The Netherlands
i.cace@ati-a.nl

² Intelligent Systems Group, ICS
Utrecht University
Princetonplein 5
3584 CC Utrecht
The Netherlands
j.j.c.meyer@uu.nl

³ University Medical Center Utrecht
Endocrinology
Huispostnummer L.00.407
3508 GA Utrecht
The Netherlands
c.r.c.pieterman@umcutrecht.nl, g.d.valk@umcutrecht.nl

Abstract: In this paper we show how knowledge from the medical domain can be incorporated in classification in a way that improves the transparency of classification (the ‘why’), and makes the classification less dependent on both the particular data set used for training and on peculiarities of the classification algorithm. We compare a decision tree incorporating a domain model with a tree built directly from the same data.

1 Introduction

Despite the many powerful classification algorithms available, AI and machine learning techniques are not part of everyday medical diagnostic practice. The usefulness of an algorithm for classifying actual medical cases, as part of a medical Decision Support System (DSS), depends on how reliable the classification is as well as on how much the practitioner trusts the algorithm’s output. In this paper we show how the know-how and concepts from the medical domain can be incorporated in a classification algorithm, based on the idea of *encapsulation* from the object oriented (OO) programming paradigm. Encapsulation is the practice of bundling methods and variables that ‘belong together’ in one object,

for example because they pertain to a specific task or data. It makes software easier to develop, test and maintain [Par72, HL95]. The effect of this approach is twofold: the algorithm should be a better model for the domain and it should make it easier to explain the classification to the user.

We will compare a classification tree incorporating domain knowledge with a classification tree built directly from data. To provide a context for evaluating classifier properties, in Section 2 we discuss reasoning in general in the medical domain, by computers and by practitioners. Then we move to the specific case of diagnosing tumors of the parathyroid glands, in Section 3. Here we describe a classification tree built directly from data, and designed to predict which patients will have surgery within a year [HPV⁺11]. In Section 4 we introduce a classification tree that we built using the same algorithm and the same data-set, but that incorporates domain knowledge. Finally in 5 we compare the two classification trees and discuss our approach in the context of medical reasoning and the current role played by medical DSSs.

2 Machine Learning Models versus Expert Models

In order to compare the way medical practitioners usually reason and the way an artificial system reasons, both need to be described in terms of the same concepts. Let a ‘model’ denote the description of the problem domain in terms relevant to that domain. For a rule-based DSS the model would be the knowledge base and the inference engine. For a single machine learning (ML) technique the model is the algorithm after learning, for example a classification tree. For a medical practitioner the model consists of all the know-how the practitioner has acquired through training and through practice that applies to the question of how to proceed given a patient’s medical findings. Note that in particular cases doctors base their professional opinions largely on models much more restricted than ‘all their medical know-how’. These include guidelines, protocols, reference ranges, and so on.

2.1 Machine Learning Models

Machine *learning* means building a representation of some part of the problem domain. The resulting model is constrained by the algorithm; if the algorithm is poorly chosen for a domain, the model can never be really good. For example a linear classifier will not do if the data is not linearly separable. It is a well-known fact in the machine learning community that there is no single learning algorithm that performs best on all data sets [Aha92]. The parameters of a model are obtained by learning from one or more data sets or by modeling domain knowledge elicited from an expert. In the first case the model is as reliable and as generalizable as the data set. This reliance on data has the benefit that the base of the model is traceable. The classification process may consist of several steps which may themselves be ML or statistical methods (for example finding the most predictive variables based on Information Theory), or may be based on domain knowledge

(for example picking predictive variables based on expert opinion).

2.2 Medical Practitioners' Models

The models practitioners have at their disposal vary greatly to the extent of how formal (or formalizable) they are. Reference ranges for example, can be expressed as formal rules: $a \leq x \leq b$. The practitioners' general know-how, medical and common-sense is at the other end of the spectrum. This type of knowledge constitutes a large part of the doctors' medical model and presumably helps doctors pick the more constrained model to apply to a particular patient's situation. The basis of medical models is formed by consensus among experts. Sometimes, in modern, evidence-based medicine, consensus can be traced back to meta-analysis of many data sets (e.g. [PML⁺01]).

3 Models of hyperparathyroidism in MEN1 patients

In this section a comparison is made between different models that apply to diagnosing a specific medical condition: parathyroid tumors in patients with Multiple Endocrine Neoplasia Type 1 (MEN1). Such tumors cause an increase in parathyroid hormone (PTH) secretion called primary hyperparathyroidism (PHPT). The affected glands have to be surgically removed to prevent disturbances in calcium metabolism and other processes [PVD⁺11].

3.1 Medical Models of PHPT

When presented with a patient with suspected PHPT, the clinician will look for deviations in calcium and PTH levels, in addition to assessing other possible predictors. Two different medical models apply to calcium and PTH levels as predictors of PHPT: the normal ranges and the sigmoid relationship between calcium and PTH (see graph in Fig. 1). PHPT patients show the same sigmoid curve, only displaced upwards and to the right. The exact shape of the curve as well as the location of the top and bottom asymptotes (the two horizontal lines) differs across subjects. The individual differences are larger for PHPT patients than in healthy subjects, and the distance between the two asymptotes is also generally larger for PHPT patients [SM94, Par01]. The following observations can be made about the two models: a) Normal ranges and the sigmoid relationship can both be expressed formally; b) There is a discrepancy between the models: the normal ranges describe a rectangular area in the Calcium-PTH plane, disregarding the sigmoid relationship between calcium and PTH; c) The discrepancy between the two models does not confuse practitioners, who combine the two to determine if a patient's values are abnormal; and d) The sigmoid curve describes the *short term* relationship between calcium and PTH, the changes that occur in PTH levels, minutes after changes in calcium levels have been arti-

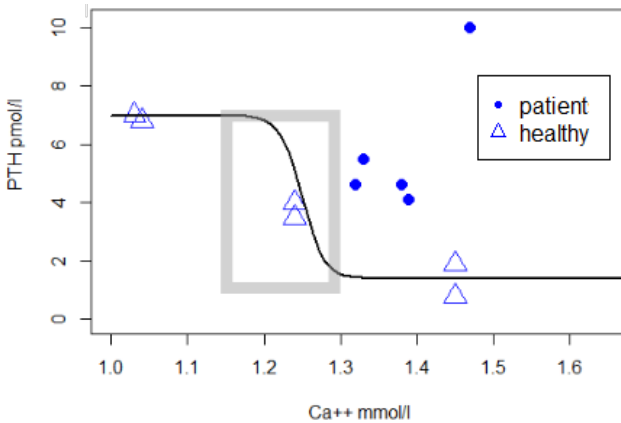


Figure 1: Sigmoid curve of the calcium-PTH relationship. Normal values are inside the gray rectangle (exact reference ranges differ per laboratory.) Also shown: values found in a healthy subject when calcium levels were manipulated [SHC⁺10], and in a MEN1 patient prior to surgery.

ficially induced. Without manipulation, values along the slope of the sigmoid (and inside the reference ranges) are found.

3.2 A Machine Learning Model of PHPT in MEN1

Höhle et al. trained [HPV⁺11] a decision tree generator [Kot07, Qui86] to find MEN1 patients who needed surgery based on yearly average calcium and PTH levels. They used discretized, average values for three years in a row. The data came from the MEN1 data set collected in Utrecht and Nijmegen. The best trees perform well: 91% of the training cases and 95% of the test cases from the dataset were classified in concordance with the surgeon's decision.

The first junction in the best tree occurs at a calcium level of 1.3 mmol/l, corresponding to the upper limit of the normal range. Higher average calcium levels indicate surgery but only in combination with high PTH. These results illustrate the importance of incorporating domain knowledge in classification. If there had not been a category boundary at or near 1.3 mmol/l, a decision threshold coinciding with the medical guidelines (i.e. normal range) simply would not be found. Calcium and PTH were discretized with the medical models in mind [HPV⁺11]. The MEN1 data show (Fig.2) that the values of both calcium and PTH are higher as patients are closer to surgery, as could be expected from the medical model discussed earlier. The ideal boundary to separate patients for surgery would be a downward slope, more or less parallel to the slope of the sigmoid curve, see Fig.3. Decision trees are generally not the best classifier to represent this type of model. To approximate a slope a tree would need to incorporate more splits on the same variables

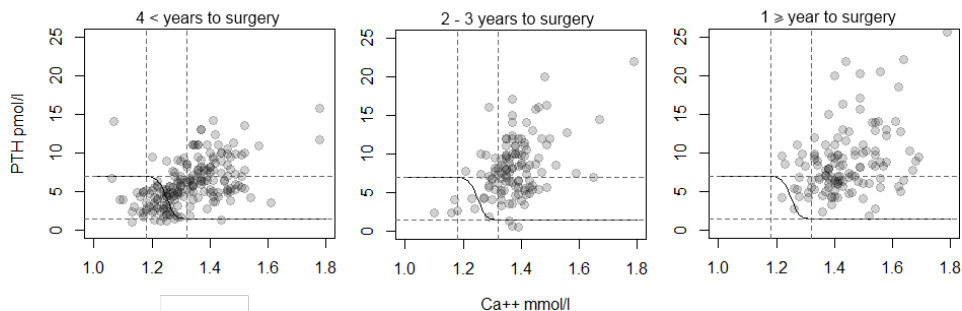


Figure 2: Calcium and PTH for MEN1 patients at different times prior to surgery. Dashed lines added for better comparison between the plots.

as illustrated in Fig.3. This is problematic because in many decision tree implementations each split would have to be coded as a separate junction, while at the same time the number of junctions is limited to prevent over-fitting [KQ02]. ML techniques that are well suited to represent sloping boundaries exist, for example Support Vector Machines (SVM) [CV95]. In this particular case, given the prior knowledge of the medical model, we would like to restrict the classifier models to *only* those incorporating the sloping boundary. We do this in the next section.

4 Encapsulated Models for Machine Learning

We will now show how the relationship between calcium, PTH and time-to-surgery can be encapsulated in a model that, given calcium and PTH levels, provides a single measure of how much the values deviate from the expected. This new measure should correspond with the severity of the PHPT, and serves as input in further classification steps. From a computational point of view it reduces the data by one dimension. We used the newly defined measure to find patients who would have surgery within a year, and compared the results with those of the classification tree described in the previous paragraph.

The remainder of this section describes classifier design, consisting of the following steps: data preparation, model definition and parameter fitting, discretization, and tree building.

4.1 Data Preparation

The MEN1 data was augmented with a binary variable ‘year’, indicating patient measures taken one year prior to surgery. When available, average values of calcium and PTH for 1 to 3 years prior were calculated. As there is no way of knowing how ‘far off’ surgery is at the time of a patients’ last record, records were cut-off either one year prior to the last available measurement or after the last surgery (whichever cut-off point was last). Records

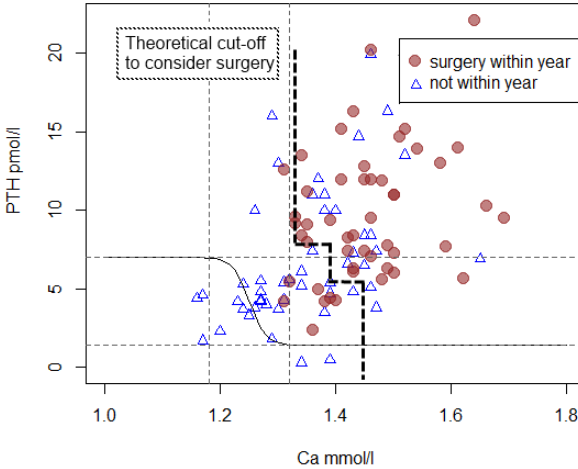


Figure 3: A univariate decision tree would need to make multiple splits on both calcium and PTH variables to encode for the changes in PTH and calcium in patients closer to surgery.

missing either calcium or PTH values were removed.

4.2 Encapsulating PHPT

The medical model of the sigmoid relationship between calcium and PTH, as described in section 3.1, is taken as the starting point. Sigmoid functions are described with 4 parameters. Function $f(x)$ (Eq. 1) is a sigmoid function, while $g(x, d)$ (Eq. 2) is that sigmoid function moved upward and to the right, at some distance d from $f(x)$. It is easy to see that for $d = 0$, $g(x, d)$ reverts to $f(x)$. Table 1 lists the parameters of $f(x)$ and $g(x, d)$.

$$f(x) = \frac{A - B}{1 + e^{r(x-c)}} + B \tag{1}$$

$$g(x, d) = \frac{A - B + d(1 - \varphi)}{1 + e^{r(x-(c+\rho d))}} + B + (\varphi d) \tag{2}$$

As the values of calcium and PTH found in clinical practice are those located along the slope, the relationship could have been modeled linearly, see Figure 4. By modeling the relationship as a sigmoid, the slope is modeled as a function d . Also the sigmoid curve corresponds with the relationship between calcium and PTH doctors are familiar with, and it has a (marginally) higher Area Under Receiver Operating Characteristic Curve (AUC) as a predictor for ‘year’.

name	description	tested values	best
A	top asymptote	8, 15	15
B	bottom asymptote	0.8, 1, 1.4	0.8
c	x value at set-point	1.225, 1.25, 1.28	1.225
r	controls the slope, the 'width' of the sigmoid graph	40, 55, 75, 100, 130	55
φ	controls the ratio between increase in upper and lower asymptote	0.1, 0.2, 0.4, 0.8, 1, 1.2, 1.5, 1.8	0.1
ρ	controls the displacement of the set-point	0.004, 0.008, 0.016, 0.032, 0.04, 0.08, 0.1	0.016

Table 1: Parameters defining the sigmoid function and the sigmoid model.

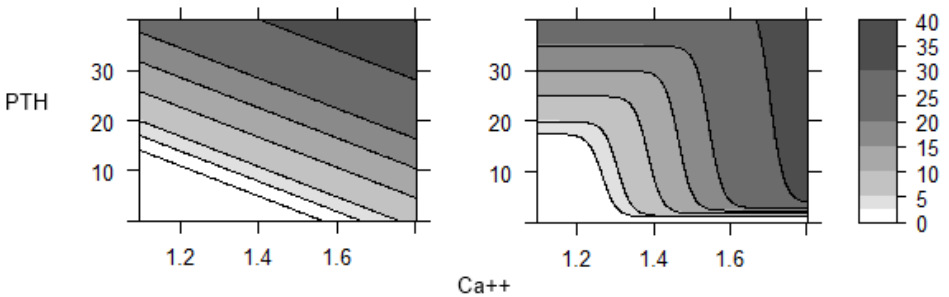


Figure 4: A linear model of calcium and PTH as a predictor for time-to-surgery, compared to the sigmoid model.

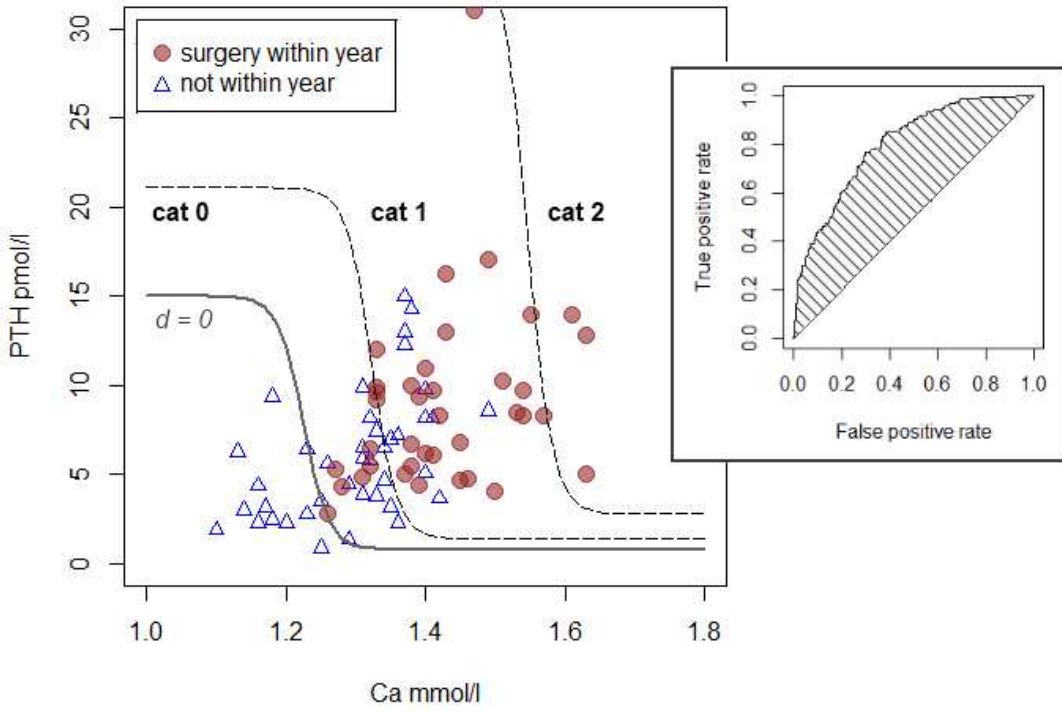


Figure 5: Graph showing the categories of d and a sample of calcium and PTH from the MEN1 data set. Inlay shows the Receiver Operating Curve for d as a predictor for surgery-within-the-year.

4.2.1 Parameter Fitting for the Sigmoid Function

Because there is no straightforward objective function, a good set of parameters was picked using a brute force approach. Specifically, parameter ranges are restricted taking into consideration the literature discussed in section 3.1 and all combinations of the values given in Table 1 were tested. We measured how well d predicted time to surgery using AUC (computed in R with the ROCR [SSBL05] package) as a predictor for ‘year’. Further investigation revealed that most of the parameters are constant for all best scoring parameter combinations, that is, they have the same or similar values. Only r and ρ varied. Bootstrapping was used to validate the findings. The parameters with the highest AUC, give us the following $f(x)$ and $g(x, d)$:

$$f(x) = \frac{14.2}{1 + e^{55(x-1.225)}} + 1 \tag{3}$$

$$g(x, d) = \frac{14.2 + 0.9d}{1 + e^{55(x-(1.225+0.016d))}} + 1 + (0.1d) \tag{4}$$

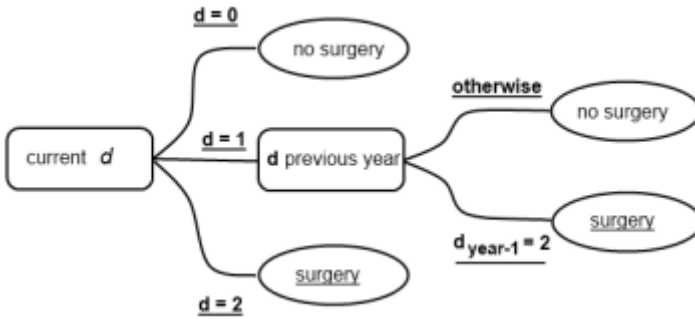


Figure 6: Decision tree for identifying patients that will be due for surgery within a year.

Combined calcium and PTH levels are only indicative of PHPT when finds are located above the graph for $f(x)$. We defined $d = 0$ for levels beneath $f(x)$ and used $g(x, d)$ to find d in all other cases. Function $g(x, d)$ returns the PTH level for the input calcium level and d . Rewriting it to return d given calcium and PTH is cumbersome. Instead d is approximated iteratively, akin to the Bisection Method for finding roots.

4.3 Discretization and Tree Building

The tree-building algorithm requires that the input variables be discrete. The newly calculated value d was discretized according to the Minimum Description Length principle [FI93] using Orange Canvas software [CDX⁺05]. This resulted in three categories, with cut-points at 6.1 and 20.0. The following input variables were supplied to the Moku tree building algorithm [HPV⁺11]: $d_{cat.}$ calculated from the input values of calcium and PTH; at most 3 historical values of d (also categorized) calculated from the average values of calcium and PTH over the previous, one, two and three years. These historical values were not always available.

4.4 Evaluation

The decision tree in Figure 6 has an accuracy of 90%, on both the test and trainings-set. Once the d has been computed and the category is known, the decision process becomes very simple as there are only three categories and two junctions in the tree. Together these cover all possible finds. The tree has a Positive Predictive Value of 0.91 and a Negative Predictive Value of 0.65. This is understandable as the data-set is highly skewed, with 969 negative examples and 126 positive examples.

5 Discussion

We used a generic machine learning technique, tree building, combined with a model provided for us by medical science, to solve a classification task. The tree is based on the input data, and classification performance measures how well the tree represents any patterns found in the input data. However, the input data may not provide a sufficient and accurate description of the problem. For example, we only used two measurements, PTH and calcium, to predict the outcome. In real life a doctor will consider more factors before reaching a decision. Also, it is unknown if the doctor's decision was correct.

The decision tree we found using the Moku tree building algorithm is small, simple and codes for all possible finds. Compared to the original tree found by Höhle et al. it is slightly less accurate: they achieve 91% accuracy for the training cases and 95% for the test cases, we have 90% accuracy on both test and train cases. The original tree was trained directly from (categorized) data while our tree used the variable d that encapsulates the relationship between calcium and PTH in a single measure that indicates 'far off' from the normal values the patient findings were. In other words, by merging calcium and PTH into one variable we enforced that the classifier observes the connexion between the two. The original classifier finds this connexion 'on its own', but it also allows for exceptions. For example, surgery is indicated in cases where both last years calcium and calcium three years ago were high, if the accompanying PTH of last year was also elevated. Except when last years' PTH was *highly* elevated (i.e. in the highest category). It is not clear whether this represents a valid reasoning pattern or a case of over-fitting.

Ideally a measure based on a medical model should be less prone to over-fitting than a measure obtained from an all purpose machine learning algorithm trained on a single data-set. MEN1 is a rare hereditary condition. We only had a small set of data collected at two cities in the Netherlands, containing a little over 1000 records for 117 patients. So the danger of over-fitting is real. To test if incorporating medical knowledge produced a more general classifier would require an additional, different and larger, data-set.

The difference in procedure between the approach described in Section 4 and the one taken by Höhle et al. is not that big. In both cases the input data was prepared with domain knowledge in mind, prior to feeding it to the tree building algorithm. Data preparation, in some form or other, is generally the first step before applying ML techniques to a data set. And it usually requires some background knowledge.

We tried to be thorough in applying medical background knowledge in the data preparation phase. Calculating d from calcium and PTH is a constituent part of the classifier, and it is described explicitly so that it can be evaluated and if necessary improved. On the down-side, the more a classifier depends on domain knowledge (as opposed to learning from data), the more time-consuming and specific to the particular domain, building that classifier becomes. This is illustrated by studies like this example of Bayesian Network design [vdGRW⁺02].

The algorithm used to find d is not a black box. This deviation measure is something doctors can understand, and therefore they can evaluate and adjust the algorithm for d to new medical insights. As it is, the measure d only takes as input calcium and PTH

levels but other variables could be included too. For example serum PTH raises with age [MMY84] so if d is to denote the deviation from normal values it should return different d s for patients of different ages. While we did find the expected relationship between age and PTH in our data-set, attempts to use age when calculating d did not yield a better predictor. That is, we could not establish that doctors used deviation from *age*-normal PTH levels as an indicator for PHPT.

Some predictive tools that are actually used in daily clinical practice were designed using a similar approach to the one described in this paper. For example, established tools for risk estimation for prostate cancer patients use a small number of variables to calculate patient risk[LS09]. The tools use a measure called ‘Gleason score’, which is a grade representing a pathologist’s professional judgment of tumor tissue, in combination with a few other patient findings. The latter are either continuous variables, from which individual risk is calculated using an algorithm equivalent to a Naive Bayes classifier. Or, the findings are categorized and summed to find which risk category the patient fits into. In other words, the input variables are patient findings translated into meaningful units of information corresponding to severity categories. This is a form of encapsulation, because the raw data and the method of deriving the information from the data is hidden from the rest of the classifier, which only deals with the end value. Another real-life example is the first trimester screening for fetal chromosomal abnormalities [KHdMS12]. Screening is routinely offered to pregnant women at risk and is fully integrated in the software used by obstetricians. Here too patient findings are translated into more meaningful units of information, which are then fed into a Naive Bayes classifier. Specifically, the classifier uses weight corrected, Multiples of the Median (MoM, $x_{MoM} = x/\mu$) of two proteins measured in patient’s blood. Note that given a patient measurement, the MoM value depends on which population is taken to calculate the median: the MoM value encapsulates the individual patient’s place in the appropriate (sub-)population.

The classification process described in this paper consists of two steps that are designed to align with the clinician’s cognitive model of the problem domain. The first step of calculating d corresponds to judging how ‘far off’ the calcium and PTH findings are. Then the next classification step takes into account current and previous findings before reaching a conclusion. The type of predictive tools that are in use by practitioners indicate that if the measure d were to be verified and trusted by domain experts (and not an ad-hoc model devised by a computer scientist), a classifier incorporating this measure could be an acceptable decision support tool. Like the variables used in the predictive tools discussed in the previous paragraph, d is independent of the rest of the classifier and can be evaluated and improved separately, analogous to Object design in OO software engineering. In general, combining different algorithms in one decision support tool does not hinder the explicability of the decision process. It may even facilitate it. There is evidence from cognitive science that as medical practitioners gain expertise, they reason less from data directly, instead they encapsulate the data into clinically relevant concepts and use these in their diagnostic reasoning [RSB00].

6 Conclusion

We compared two decision trees built from the same data-set, but in the future the same should be repeated for other ML techniques. Preferably this should be done with a more extensive data-set. In this case we only had 3 variables to work with: calcium, PTH and age. Other variables like patient kinship and genetic mutations contained too many values to be useful.

In this paper we used 3 fairly simple techniques to build a classifier: we fitted a mathematical model to data, discretized the output and built a decision tree with the result. But nothing stands in the way of leveraging the full power of ML and Artificial Intelligence in general to deliver high quality decision support. That is, by carefully choosing algorithms and methods according to the sub-problems to model, it is possible to pick an algorithm that is well suited to model each particular sub-problem while at the same time preventing over-fitting to a particular data-set, by constraining the search to models that comply with expert opinion of the domain. This also means the algorithm incorporates concepts users are familiar with, making the reasoning process more transparent and aiding the acceptance of the algorithm as a tool for decision support.

7 Acknowledgments

The authors wish to thank Dr. Ameen Abu-Hanna for helping us get a better picture of what is expected of predictive tools in the medical domain and for his pointers on statistical methods.

References

- [Aha92] David W. Aha. Generalizing from Case Studies: A Case Study. In *Proceedings of the Ninth International Conference on Machine Learning*, pages 1–10. Morgan Kaufmann, 1992.
- [CDX⁺05] Tomaž Curk, Janez Demšar, Qikai Xu, Gregor Leban, Uroš Petrovič, Ivan Bratko, Gad Shaulsky, and Blaž Zupan. Microarray data mining with visual programming. *Bioinformatics*, 21:396–398, February 2005.
- [CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
- [FI93] Usama M. Fayyad and Keki B. Irani. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In *IJCAI*, pages 1022–1029, 1993.
- [HL95] Walter L. Hürsch and Cristina Videira Lopes. Separation of Concerns. Technical report, College of Computer Science, Northeastern University, 1995.
- [HPV⁺11] D. Höhle, C. R. C. Pieterman, G. D. Valk, A. R. Hermus, H. P. F. Koppeschaar, John-Jules Ch. Meyer, and R. P. J. de Lange. Classifying the decision to perform surgery

in MEN1 cancer patients using decision trees. *Computer-Based Medical Systems, IEEE Symposium on*, pages 1–6, 2011.

- [KHdMS12] Maria PH Koster, Kirsten M Heetkamp, Esteriek de Miranda, and Peter CJI Schielen. Comparison of risk calculation approaches in a screening programme for Down syndrome. *Journal of perinatal medicine*, 40(3):259–263, 2012.
- [Kot07] Sotiris B. Kotsiantis. Supervised Machine Learning: A Review of Classification Techniques. In *Emerging Artificial Intelligence Applications in Computer Engineering*, pages 3–24. IOS Press, 2007.
- [KQ02] Ronny Kohavi and J. Ross Quinlan. Handbook of data mining and knowledge discovery. chapter Data mining tasks and methods: Classification: decision-tree discovery, pages 267–276. Oxford University Press, Inc., New York, NY, USA, 2002.
- [LS09] William T. Lowrance and Peter T. Scardino. Predictive Models for Newly Diagnosed Prostate Cancer Patients. *Reviews in Urology*, 11:117–126, 2009.
- [MMY84] R. Marcus, P. Madvig, and G. Young. Age-related changes in parathyroid hormone and parathyroid hormone action in normal humans. *The Journal of clinical endocrinology and metabolism*, 58(2):223–230, 1984.
- [Par72] David Lorge Parnas. On the Criteria To Be Used in Decomposing Systems into Modules. *Commun. ACM*, 15(12):1053–1058, 1972.
- [Par01] A. Michael Parfitt. Parathyroid Growth: Normal and Abnormal. In John P. Bilezikian, Robert Marcus, and Michael A. Levine, editors, *The Parathyroids: Basic and Clinical Concepts*, chapter 18, pages 293–330. Academic Press, 2001.
- [PML⁺01] Alan W Partin, Leslie A Mangold, Dana M Lamm, Patrick C Walsh, Jonathan I Epstein, and Jay D Pearson. Contemporary update of prostate cancer staging nomograms (Partin Tables) for the new millennium. *Urology*, 58(6):843 – 848, 2001.
- [PVD⁺11] C.R. Pieterman, M.R. Vriens, K.M. Dreijerink, R.B. van der Luijt, and G.D. Valk. Care for patients with multiple endocrine neoplasia type I: the current evidence base. *Familial Cancer*, 10(1):157–71, March 2011.
- [Qui86] J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.
- [RSB00] Remy M. J. P. Rikers, Henk G. Schmidt, and Henny P. A. Boshuizen. Knowledge Encapsulation and the Intermediate Effect. *Contemporary Educational Psychology*, 25:150–166, 2000.
- [SHC⁺10] Rajiv P. Shrestha, Christopher V. Hollot, Stuart R. Chipkin, Claus P. Schmitt, and Yossi Chait. A mathematical model of parathyroid hormone response to acute changes in plasma ionized calcium concentration in humans. *Mathematical Biosciences*, 226:46–57, 2010.
- [SM94] Peter Schwarz and Jesper C. Madsen. Parathyroid hormone dynamics in hypoparathyroidism and primary hyperparathyroidism. *Journal of Bone and Mineral Metabolism*, 12:S75–S79, 1994.
- [SSBL05] Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. ROCRC: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941, Oct 2005.
- [vdGRW⁺02] Linda C. van der Gaag, Silja Renooij, C. L. M. Witteman, B. M. P. Aleman, and B. G. Taal. Probabilities for a probabilistic network: a case study in oesophageal cancer. *Artificial Intelligence in Medicine*, 25(2):123–148, 2002.