Acquisitie van medische kennis ten behoeve van expertsystemen

Redactie:
E. Backer
J.H.C. Reiber
J.W. Smeets

Delftse
Universitaire
Pers

Acquisitie van medische kennis ten behoeve van expertsystemen

Acquisitie van medische kennis ten behoeve van expertsystemen

Redactie:
E. Backer
J.H.C. Reiber
J.W. Smeets

Delftse Universitaire Pers/ 1990

Uitgegeven en gedistribueerd door:

Delftse Universitaire Pers
Stevinweg 1
2628 CN Delft
Tel. 015-783254

In opdracht van:

Technische Universiteit Delft, Vakgroep Informatietheorie
Erasmus Universiteit Rotterdam, Thoraxcentrum
Stichting Centrum Medische Techniek (Tel. 01802-2089)

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Acquisitie

Acquisitie van medische kennis ten behoeve van expertsystemen / red.: E. Backer, J.H.C. Reiber, J.W. Smeets. - Delft : Delftse Universitaire Pers. - Ill.
Uitg. in opdracht van: Stichting Centrum Medische Techniek, Zevenhuizen. - Met lit. opg.
ISBN 90-6275-607-7
SISO 527.8 UDC 681.324:61 NUGI 743
Trefw.: expertsystemen : medische techniek.

Inhoudsopgave

VOORWOORD

Van expertsystemen wordt verwacht dat ze kunnen redeneren zoals de menselijke experts dat doen. Dit redeneren geschiedt meestal volgens de door de kennisingenieur opgestelde regels. Echter, om deze regels te kunnen opstellen dient de kennisingenieur allereerst te begrijpen volgens welke "ervarings"-regels de expert redeneert. De kwaliteit van dit proces van "kennisacquisitie" bepaalt vanzelfsprekend voor een groot gedeelte de uiteindelijke kwaliteit van het te realiseren expertsysteem.

In de praktijk blijkt dat dit verzamelen van kennis een zeer moeilijke proces is; veelal kunnen de experts zelf niet uitleggen waarom ze tot een bepaalde conclusie komen. Bovendien blijkt de kennis veelal een zekere mate van onzekerheid te bevatten.

In dit boek zullen verschillende aspecten van kennisacquisitie worden toegelicht. "Onzekerheid" speelt een belangrijke rol in de interpretatie door de experts. Hoe deze onzekerheid kan worden gerepresenteerd en gemanipuleerd, wordt beschreven door Backer. De invloed van de aanwezige biomedische en klinische kennis van de experts wordt besproken in het hoofdstuk van Boshuizen. Krijgsman beschrijft aan de hand van een praktische situatie op welke wijze en in welke mate kennis onttrokken kan worden van de experts, in zijn geval op het gebied van de thallium-201 tomografie. In de bijdrage van Maes wordt aangegeven op welke wijze anatomische kennis werd verworven en geïmplementeerd in een systeem voor de automatische labeling van bloedvaten in angiogrammen. Post beschrijft aan de hand van een medisch expertsysteem in hoeverre technieken voor automatisch leren toe te passen zijn. Tenslotte wordt de life-cycle van medische beslissingsondersteunende systemen beschreven door Jaspers.

De editors hopen dat dit boek mag bijdragen tot een beter begrip van de mogelijkheden en beperkingen van de huidige kennis-acquisitie-methoden en mag leiden tot de ontwikkeling van nieuwe, verbeterde technieken op dit gebied.

Delft, april 1990

E. Backer
J.H.C. Reiber
J.W. Smeets

OVER DE ACQUISITIE (EN REPRESENTATIE) VAN ONZEKERHEID IN (EN OVER) KENNIS TEN BEHOEVE VAN MEDISCHE EXPERT SYSTEMEN

E.Backer en J.C.A. van der Lubbe

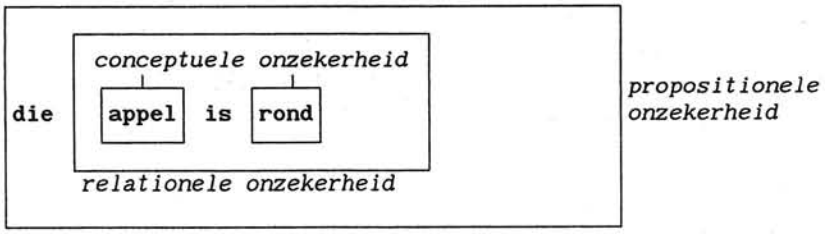
Technische Universiteit Delft
Faculteit der Elektrotechniek
Vakgroep Informatietheorie

1 Introductie

Kennisgestuurde systemen (waaronder expert systemen) zijn bedoeld om door middel van manipuleren (redeneren) van kennis en informatie bijvoorbeeld een probleem op te lossen dan wel een diagnose te stellen. Zowel kennis als aangeboden informatie kunnen niet precies, inkompleet of vaag zijn. We zullen dat aanduiden met 'onzekerheid'.Sinds de zestiger jaren trachten onderzoekers computerprogramma's te schrijven welke in staat zijn op basis van door de patient aangedragen symptomen en op basis van in de computer opgeslagen expertkennis over het probleemgebied, automatisch een diagnose te genereren. Als zodanig imiteren of representeren deze systemen een stukje 'subjectief' menselijk (expert-) redeneren, althans in termen van input-output-gedrag.
Karakteristiek voor menselijk redeneren is het vermogen te kunnen omgaan met onzekere en niet precieze informatie. Veel aandacht is derhalve geschonken aan de wijze waarop deze onzekerheid in de computer is te representeren en hoe er mee te manipuleren, zodat tenminste het input-output-gedrag enigszins overeenkomt met het subjectief menselijk redeneergedrag.
De betekenis van onzekerheid echter is verre van eenduidig. Onzekerheid in een kennispropositie kan te maken hebben met de 'geloofwaardigheid' van de propositie, met de 'statistische geldigheid' ervan en met intrinsieke 'vaagheid' voor wat betreft de in de propositie gehanteerde objecten en attributen.
In het volgende voorbeeld kunnen we de diverse klassen onzekerheid aanduiden:

in de propositie
die appel is rond
is **appel** het object en **rond** het attribuut.

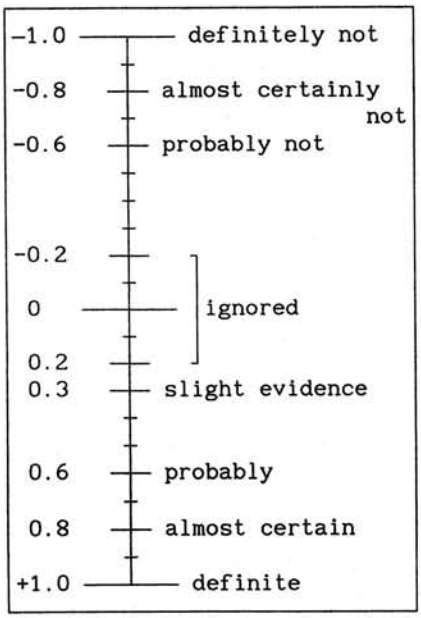
De resulterende hierarchische nesting van 'onzekerheden' voor deze propositie ziet er dan als volgt uit:



In het navolgende beperken we ons tot de koppeling van deze drie klassen van onzekerheden en drie typen van representatie. Met de *propositionele onzekerheid* is 'geloofwaardigheid' geassocieerd, de *relationele onzekerheid* wordt gerepresenteerd door statistische geldigheid of nauwkeurigheid en *conceptuele onzekerheid* wordt weergegeven door intrinsieke object- en attribuutvaagheid.

Een aantal aspecten van het representeren en manipuleren van onzekerheid in expert systemen zijn in de loop van de ontwikkeling van deze systemen dominant naar voren gekomen.

(i) *het representeren van onzekerheid*
In het algemeen beeldt men een informeel (kwalitatief) waardeoordeel over de propositie af op een numerieke schaal. Figuur 1 geeft een voorbeeld van een dergelijke afbeelding weer. Het resultaat is dat aan iedere kennisregel een getal wordt toegekend dat iets zegt over de onzekerheid in de regel of over de gebruikswaarde van die regel. Dit getal geven we aan met <cf>, de zekerheidsfactor.



Figuur 1: informele maat voor geloofwaardigheid of zekerheid

We noteren dan bijvoorbeeld

{kennisregel(objecten,attributen)}<cf> of
{kennisregel(objecten,objecten)}<cf>

(ii) *het manipuleren van onzekerheid*
Als in de objecten of attributen onzekerheid is vastgesteld dan dient een herwaardering van de propositie plaats te vinden, bijvoorbeeld als volgt.

{kennisregel(objecten<cf>$_1$,(attributen<cf>$_2$}<cf'>
waarin
<cf'> = <cf> MIN [<cf>$_1$,<cf>$_2$].

een van de mogelijkheden is om tot herwaardering te komen.

Indien kennisregels in samenhang worden beschouwd (redeneren), bijvoorbeeld:

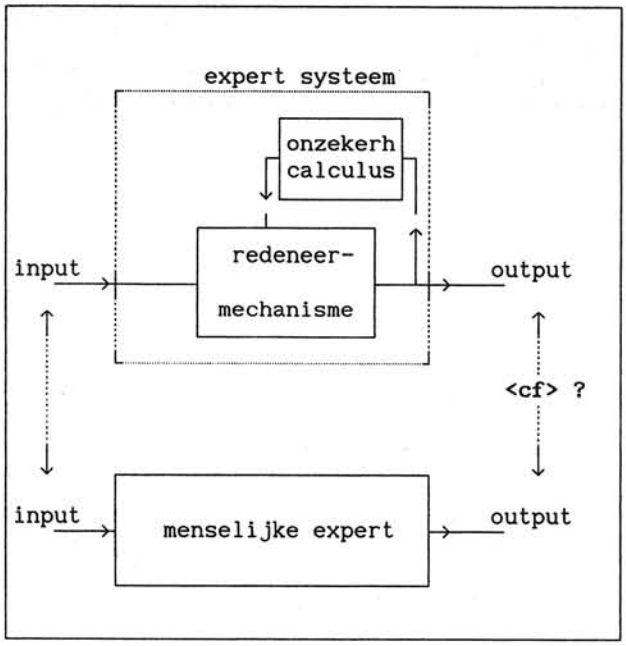
conclusie[{kennisregel i}<cf>$_i$,{kennisregel j}<cf>$_j$]<cf>

dan zal **<cf>** mogelijk een functie zijn van **<cf>$_i$** en **<cf>$_j$**.

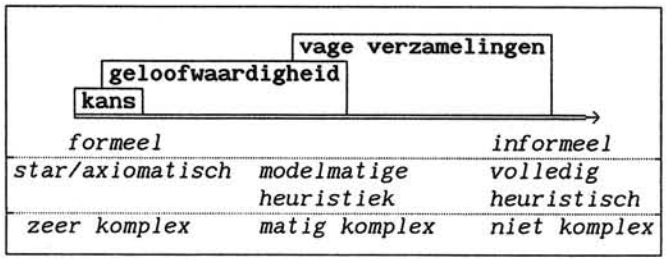
We spreken van propagatie van onzekerheid.

De wijze waarop het herwaarderen en propageren van onzekerheid is vastgelegd noemen we een onzekerheidscalculus.
Een dergelijke calculus kan gebaseerd zijn op het zuivere kansbegrip (Bayes), op geloofwaardigheidsmaten (methode Shortliffe-Buchanan en de theorie van Dempster-Shafer) of op de theorie van de vage verzamelingen (Zadeh).[1]
Ieder van deze mogelijkheden zijn gekarakteriseerd door rekenkundige komplexiteit, veronderstellingen, voor- en nadelen.
Indien we ten doel stellen dat het input-output-gedrag van een expert systeem in voldoende mate overeenkomt met het subjectief expert redeneren, dan kan de keuze van het representeren van onzekerheid en de in het expert systeem ondergebrachte onzekerheidscalculus van doorslaggevende betekenis zijn voor het bereiken van dit doel, zie figuur 2.

[1] hanteren van het zuivere kansbegrip impliceert probabilistische additiviteit $p(u) + p(\neg u) = 1$ en probabilistische implicatie en vereist numerieke kompleetheid; de overige calculi hanteren heuristische maten en beantwoorden aan de wenselijke eigenschap dat $f(u) + f(\neg u) \leq 1$. De geloofwaardigheidsmaten zijn pseudo-statistisch. De lidmaatschapswaarden van vage verzamelingen zijn volstrekt heuristisch van aard.



Figuur 2: de keuze van het redeneermechanisme en de onzekerheidscalculus is van beslissende betekenis voor de vergelijking van het input-output-gedrag van mens en systeem.



Figuur 3: ordening van onzekerheid

In figuur 3 is, als uitkomst van de vele hiermee samenhangende literatuur [2], een ordening van onzekerheidsbeginselen (het zuivere kansbegrip, de geloofwaardigheid en de conceptuele vaagheid) aangegeven.
Hieruit kunnen we direct een verdere probleemstelling destilleren. Het lijkt een over-vereenvoudiging om aan een kennispropositie slechts één onzekerheidsindikator mee te geven.

In deze bijdrage zullen we ervan uitgaan dat aan iedere kennispropositie én een geloofwaardigheid (propositionele onzekerheid), én een statistische geldigheid (relationele onzekerheid) én een intrinsieke conceptuele vaagheid (conceptuele onzekerheid) is verbonden.
Dus:

$$\{\textbf{kennisregel(objecten,attributen)}\} \begin{bmatrix} \textit{geloofwaardigheid} \\ \textit{stat. geldigheid} \\ \textit{concept. vaagheid} \end{bmatrix}$$

waarmee tevens gezegd is, dat de onzekerheidscalculus zo ingericht zal dienen te zijn dat deze vormen van onzekerheid gelijktijdig kunnen worden gemanipuleerd.
We zullen tevens aangeven dat de acquisitie van deze onzekerheid goed (beter?) aansluit bij het gebruik maken van expert-panels (als een gestruktureerd proces van kennisacquisitie).

2. De representatie in de vorm van onzekerheidsvectoren

In de praktijk ontdekken we dat er spanning bestaat bij het formuleren van de kennis door een expert tussen de 'logische struktuur' van de propositie en de gehanteerde taal als substraat voor de propositie.
Het volgende voorbeeld maakt dat duidelijk.

a. *iedere roker krijgt longziekte*
b. *rokers krijgen longziekte*
c. *rokers krijgen longziekte <cf=80>*
d. *80% van de rokers krijgen longziekte*
e. *deze roker krijgt longziekte*
f. *een roker krijgt eerder longziekte dan een niet-roker*

De uitspraken zijn opzich willekeurig gekozen. Informeel gesproken zijn ze enigermate geordend naar 'niveau van kennis' in termen van

[2] Informele introducties zijn te vinden in Harmon & King (1985), Tanimoto (1987), Rich (1983) en Luger & Stubbefield (1989); onzekerheidscalculi worden besproken en vergeleken in Buchanan & Shortliffe (1984), Shafer (1975), Prade (1985) en Henkind & Harrison (1988). Het concept van vage verzamelingen in expert systemen is te vinden in o.a. Leung & Lam (1988).

'generaliserend vermogen'. In termen van 'soort' onzekerheid zijn ze onderscheidelijk.

Uitspraak **a.** is absoluut generaliserend. Iedere vorm van statistische onzekerheid is geelimineerd. Het object 'roker' en attribuut 'longziekte' zijn mogelijk conceptueel vaag (of onzeker) maar in deze uitspraak irrelevant (pas indien we te maken hebben met "is meneer Pieterse een roker?" wordt conceptuele vaagheid relevant). In deze vorm is de uitspraak wel onderhevig aan een mate van geloofwaardigheid (als 'overstatement' is de geloofwaardigheid ervan toch kleiner dan 100%).

Uitspraak **b.** is eveneens sterk generaliserend maar biedt ruimte voor statistische interpretatie, hoewel de geloofwaardigheid van een dergelijk statistisch model zeer klein kan zijn.

Uitspraak **c.** is een algemene uitspraak waarbij (pseudo-) statistische geldigheid expliciet gesuggereerd wordt. De geloofwaardigheid van een statistische interpretatie is ook in dit geval klein zolang geen specifieke betekenis is toegekend aan 'roker' en 'longziekte'.

Uitspraak **d.** suggereert statistische kennis met betrekking tot het domein. De geloofwaardigheid spitst zich toe op de vraag of de expert inderdaad specifieke betekenis kan toekennen aan 'roker' en 'longziekte'.

Uitspraak **e.** is een singuliere uitspraak waarbij alle onzekerheid is teruggebracht tot een onderliggend niveau van
{een roker is ....}<cf> en
{een longziekte is ....}<cf>.

Uitspraak **f.** tenslotte is weer een algemene uitspraak waaraan men een zekere mate van geloofwaardigheid kan toekennen maar waarin de onzekerheid is terug te voeren tot de dichotomie van 'rokers' en 'niet-rokers'.
De conclusie is dat met de uitspraken **a.** en **b.** in het bijzonder "geloofwaardigheid" is geassocieerd, met de uitspraken **c.** en **d.** in hoofdzaak statistische geldigheid in het geding is en de uitspraken **e.** en **f.** hoofdzakelijk terug te voeren zijn tot conceptuele onzekerheid (intrinsieke vaagheid in het object 'roker' en attribuut 'longziekte'. Vrijwel iedere uitspraak zal dus onderhevig zijn aan het stelsel onzekerheden:

- *geloofwaardigheid van de inductie van de expert*
- *statistische geldigheid van de inductie*
- *conceptuele vaagheid in objecten en attributen waarop de inductie is gebaseerd*

Aan iedere propositie wordt derhalve een onzekerheidsvector toegevoegd met als kentallen de geloofwaardigheid, de statistische geldigheid en de conceptuele vaagheid. Figuur 4 toont de resulterende onzekerheidsruimte waarop de onzekerheidscalculus zal moeten zijn gebaseerd.

3. Het bepalen van de onzekerheidsvector

Experts kunnen zeer verschillen in de wijze van redeneren en het expliciteren van de geloofwaardigheid en geldigheid van hun kennis. Voor het conceptualiseren van redeneerpaden in een redeneermechanisme is het gebruik maken van diverse experts in het probleemgebied vaak lastig en soms hinderlijk (moeilijk tot concensus van redeneren te brengen; het proces convergeert langzaam). Voor het verkrijgen van inzicht in geloofwaardigheid en geldigheid van gegeven proposities is het gebruik maken van expert-panels een 'must'. Het is in veel gevallen zelfs denkbaar dat binnen één probleemstelling verschillende expert-panels moeten worden aangesproken.

In ons voorbeeld is in ieder geval diagnostische én conceptuele expertise van belang, dat wil zeggen dat we onafhankelijke expert-panels nodig hebben om -bij gegeven proposities- de geloofwaardigheid, de objectvaagheid en de attribuutvaagheid afzonderlijk te schatten. Deze samenhang is in figuur 5 geillustreerd.

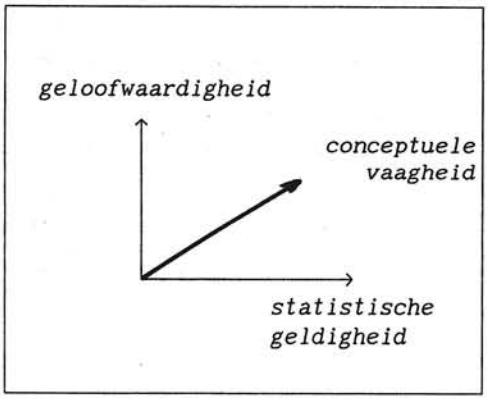
Onzekerheid in relatie tot een inductie van experts is vrijwel alleen numeriek te maken door (informele) pseudo-statistiek over groepen van experts (panels). Het betreft -onafhankelijk van elkaar- het schatten van de geloofwaardigheid van een bepaalde propositie, het bepalen van de lidmaatschapfuncties van de vage verzamelingen 'roker' (μ_{roker}) en 'longziekte' ($\mu_{\text{longziekte}}$). Bij nadere specificering van 'roker' en 'longziekte' kan volgens de statistische geldigheid worden geschat.

Recent onderzoek [3] heeft wegen geopend om deze schattingen geschikt te verkrijgen met gebruik maken van expert panels.

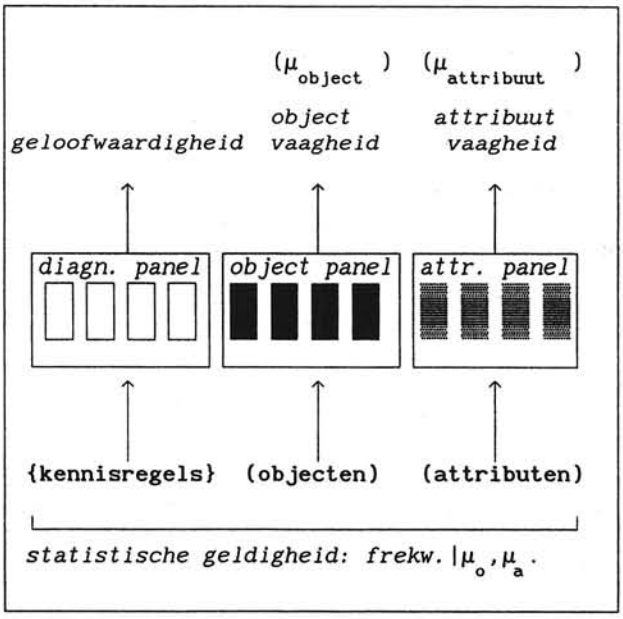
[3] In een samenwerkingsprojekt met Unilever Research Laboratory Vlaardingen zijn twee studies in de open literatuur verschenen resp. voor publicatie aangeboden:
- Fuzzy set theory applied to product classification by a sensory panel,
- The use and measurement of fuzzy logic membership functions using sensory panels; a case study.

4. Het manipuleren van onzekerheidsvectoren

Gegeven het feit dat we hier (drie) soorten onzekerheden wensen te onderscheiden dienen zich tenminste twee mogelijkheden tot manipuleren ervan aan.



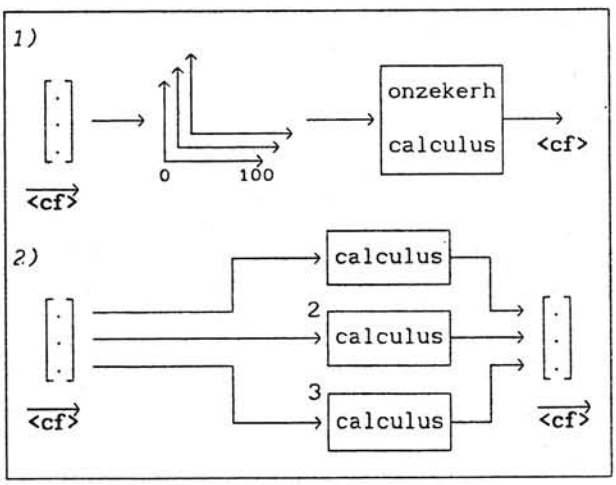
Figuur 4: de onzekerheidsruimte



Figuur 5: schatting van geloofwaardigheid, conceptuele vaagheden en statistische geldigheid

mogelijkheid 1:
We beelden ieder type onzekerheid (met een eigen afbeeldingsrelatie) af op één numerieke schaal (bijvoorbeeld [0,100]). Voor het herwaarderen en propageren is dan slechts één (geschikt te kiezen) onzekerheidscalculus vereist. Deze mogelijkheid vereist dan wel één extra (moeilijk realiseerbare) acquisitiestap, namelijk het bepalen van de onderscheidelijke afbeeldingsrelaties. Het rekenkundig voordeel wordt ruimschoots overschaduwd door de toenemende komplexiteit van de acquisitie. Het grote nadeel bovendien is dat in de uiteindelijke conclusie geen inzicht meer bestaat in de mate waarin de afzonderlijke onzekerheden hebben bijgedragen. Bovendien hebben we één calculi moeten kiezen welke voor één type onzekerheid zeer geschikt kan zijn maar voor de andere typen mogelijk veel minder geschikt.

mogelijkheid 2:
Manipulatie van onzekerheidsvectoren geschiedt met net zoveel onderscheidelijke onzekerheidscalculi als er onzekerheidskentallen in de vector zijn ondergebracht. Iedere einduitspraak (na redeneren) is dan nog steeds voorzien van van de mate waarin geloofwaardigheid , statistische geldigheid en conceptuele vaagheid aan de orde waren. In figuur 6 zijn deze twee mogelijkheden schematisch vergeleken.



Figuur 6: onzekerheidsmodellen

Recent onderzoek [4] leert dat het onzekerheidsvectormodel tot een

[4] Backer etal (1988): Modelling uncertainty in ESATS by classification inference; samenwerkingsprojekt met het Laboratorium voor Klinische en Experimentele Beeldverwerking, Thoraxcentrum, Erasmus Universiteit.

aantrekkelijke methode leidt welke op inzichtelijke wijze aansluiting geeft op zowel regel-inferentie als classificatie-inferentie bij de interpretatie van Thallium Scintigrammen.

5. Conclusie en samenvatting

In deze bijdrage hebben we ons geconcentreerd op de vraag inhoeverre het wenselijk en mogelijk is diverse typen onzekerheden afzonderlijk te bepalen en vectorieel te herwaarderen en te propageren bij combineren van meerdere proposities. Dit leidde er toe dat overeenkomstig de kentallen in de onzekerheidsvector een geloofwaardigheidscalculus, een statistische calculus en een vaagheidscalculus parallel worden aangestuurd. De acquisitie van voornoemde typen onzekerheden vereisen een aantal onafhankelijke expert-panels.
Een implementatie van het onzekerheidsvectormodel in relatie tot een classificatie-inferentiemechanisme voor ESATS (Expert Systeem voor de Analyse van Thallium Scintigrammen) wordt thans uitgevoerd.

literatuur

[1] P.Harmon & D.King (1985): *Expert Systems*. John Wiley.
[2] S.Tanimoto (1987): *The Elements of Artificial Intelligence*. Computer Science Press.
[3] E.Rich (1983): *Artificial Intelligence*. McGraw-Hill.
[4] G.F.Luger & W.A.Stubblefield (1989): *Artificial Intelligence and the Design of Expert Systems*.
[5] B.G.Buchanan & E.H.Shortliffe (1984): *Rule-based Expert Systems*. Massachusetts.
[6] H.Prade (1985): A computational approach to approximate and plausible reasoning with applications to Expert Systems. *IEEE Trans on PAMI, 7,3*.
[7] S.J.Henkind & H.Harrison (1988): An Analysis of four Uncertainty Calculi. *IEEE Trans SMC, 18,5*.
[8] E.Backer, J.J.Gerbrands, J.H.C.Reiber, A.E.M.Reijs, W.Krijgsman & H.J. vd Herik (1988): Modelling uncertainty in ESATS by Classification Inference. *Pattern Recognition Letters, 8*.
[9] E.Backer,J.C.A. vd Lubbe & W.Krijgsman (1988): On Modelling of Uncertainty and Inexactness in Expert Systems. *Proc. 9th Symp. on Information Theory, Mierlo.*
[10] L.A.Zadeh (1988): Fuzzy Logic. *IEEE Comp.*
[11] K.S.Leung & Lam (1988): Fuzzy Concepts In Expert Systems. *IEEE Comp.*
[12] M.Togai & S.Watanabe (1988): Expert System on a Chip. *IEEE Expert*.
[13] J.Gordon & E.H.Shortliffe (1985): A Method for Managing Evidential Reasoning in a Hierarchical Hypothesis Space. *AI 26*.

[14] G.Shafer (1975): *A Mathematical Theory of Evidence.* Princeton University Press.
[15] G.Shafer & Logan (1987): Implementing Dempster's Rule for Hierarchical Evidence. *AI 33.*
[16] J.Pearl (1986): Fusion, Propagation, and Structuring in Belief Networks. *AI 29.*
[17] P.L.Bogler (1987): Shafer-Dempster Reasoning with Applications to Multisensor Target Indentification Systems. *IEEE Trans. on SMC, 17,6.*
[18] R.P.W.Duin, E.Backer,S. de Jong, H.W.Lincklaen Westenberg & J.F.A.Quadt: The Use and Measurement of Fuzzy Logic Membership Functions using Sensory Panels. *Submitted to IEEE Trans. on SMC.*
[19] H.W.Lincklaen Westenberg,S. de Jong,D.A. van Meel, J.F.A.Quadt, E.Backer & R.P.W.Duin (1989): Fuzzy Set Theory Applied to Product Classification by a Sensory Panel. *Journal of Sensory Studies, 4.*
[20] T.L.Fine (1973): *Theories of Probability.* Academic Press.
[21] W.A.Gale (1986): *Artificial Intelligence and Statistics.* Addison-Wesley.

BIOMEDICAL KNOWLEDGE AND CLINICAL EXPERTISE

H. P. A. Boshuizen & H. G. Schmidt

University of Limburg, Maatricht

As early as in the 15th century, physicians and other students of human biology tried to peer into the 'black box' of the human body. Many organs and other structures in the human body were described since that time, while after the development of the multi-lensed microscope, organ structure and physiology could be studied in more detail. Through these efforts, the secrets that were kept safe in the 'box' were discovered. Important physicians such as Boerhaave (1668-1738) proved the significance of biomedical sciences (e.g. anatomy and physiology) for the clinical sciences. Research into the structure and functioning of the human body provided an increasing insight in its normal functioning and in the way it restores disturbances of its equilibrium. These research efforts resulted in a deeper insight in the mechanisms underlying long known empirical rules of thumb became understood and, as a consequence, medicine developed from an art into a modern science. In particular since the beginning of this century, the biomedical sciences play an increasingly important role in the medical curriculum.

Notwithstanding its importance for medicine as a science, the role of biomedical knowledge in medical diagnosis and treatment in everyday practice is not at all clear. Feltovich and Barrows (1984), for instance, hypothesized that biomedical knowledge plays an integrating role in the understanding and diagnosis of a clinical case. Feltovich' and Barrows' position can be paraphrased as "comprehension, and hence the diagnosis, of a case emanates from biomedical knowledge". Their point of view is supported by other investigators in the domain of medical diagnosis (e.g. Lesgold, 1984; Kuipers and Kassirer, 1984; Kuipers, 1985; Lesgold, Rubinstein, Feltovich, Glaser, Klopfer and Wang, 1988). These authors all emphasize the role of biomedical knowledge in medical reasoning.

This perspective on diagnostic reasoning, however, is challenged by Patel, Evans and Groen (1989) and others (e.g. Schmidt, Boshuizen and Hobus, 1988). These authors suggest that medical experts predominantly use *clinical* knowledge instead of biomedical knowledge to represent and diagnose a patient problem[1]. According to these investigators, the application of biomedical knowledge is in particular characteristic for *non-expert* reasoning. More generally stated: the application of biomedical knowledge is associated with non-automatic problem solving and will be found in the diagnosis of non-routine cases. But, as Boshuizen, Schmidt and Coughlin (1987) already pointed out, there is reason to assume that this debate results from incomplete models of the role and structure of clinical and biomedical knowledge at consecutive stages of the development of medical expertise. Aim of the present paper is to attain more insight in the organization of biomedical and clinical knowledge and to investigate possible mechanisms responsible for changes in the role and organization of clinical and biomedical knowledge in the course of the development from novice to expert.

---

[1]Clinical knowledge is defined here as knowledge of attributes of sick people. It concerns itself with the ways in which a disease can manifest itself in patients; the kind of complaints one would expect given that disease; the nature and variability of the signs and symptoms and the ways in which the disease can be managed. Biomedical knowledge by contrast, concerns itself with the pathological principles, mechanisms or processes underlying the manifestations of disease. It is phrased in terms of entities such as viruses or bacteria, in terms of tissue, organs, organ systems, or bodily functions.

In order to attain these goals, an experiment was designed in which the application and availability of clinical and biomedical knowledge in clinical reasoning were investigated. Clinical and biomedical knowledge *application* were measured by analyzing the subjects' think-aloud protocols. The *availability* of biomedical knowledge was assessed from the subjects' post-hoc explanation of the biomedical process underlying the patient's signs and symptoms. Four levels of expertise were incorporated and it was expected that the overt application of biomedical knowledge would decrease with an increasing level of expertise (Boshuizen, Schmidt & Coughlin, 1988). Furthermore, two variations of the same case were used: a typical and an atypical one. According to Schmidt, Boshuizen and Hobus (1988) and to Patel, Evans and Groen (1989) this atypical case, rather than the typical variant would give rise to biomedical reasoning, because physicians can only to a lesser extent rely on automatic processing while diagnosing an atypical case.

The question of knowledge development and the relative roles of biomedical and clinical knowledge will be addressed in a three step approach. The first step is to find an answer to the question 'Does the application of biomedical knowledge in clinical reasoning decrease with an increasing level of expertise?' Should this question be answered with 'yes', as is expected, then the next question is whether this decrease in the application of biomedical knowledge is associated with a decrease in the availability of this kind of knowledge in long term memory The final step aims at a clarification of the underlying developmental mechanism.

Method

In this experiment 38 subjects participated, 28 students and ten physicians. Ten subjects were second year students, eight subjects were fourth year students. Their knowledge structure and knowledge application were assessed at the end of the second semester, hence the second year students may be assumed to have acquired all relevant biomedical knowledge, while the fourth year students will have studied the relevant biomedical and clinical subjects. Furthermore, ten fifth year subjects participated who had finished their clerkships in internal and family medicine. The expert group consisted of ten family physicians with at least four years of experience.

The subjects were presented with a case of pancreatitis. The patient was a 38 year old, unemployed male with a history of neurotic depressions and alcohol abuse. One year earlier, he had been hospitalized with abdominal complaints, and now calls the family physician with a complaint of severe, boring pain in the upper part of the abdomen. This patient suffered from a chronic alcohol-induced pancreatitis. The subjects' task in this experiment was to diagnose the case while thinking aloud. After completing the case they were asked to describe (in writing) the pathophysiological processes that in their opinion underlie the case.

The case was presented in one of two forms, a typical or an atypical case of alcohol induced pancreatitis with several complications. In the typical form, both the patient's medical background and signs and symptoms fitted with what can normally be expected in this class of patients. In the atypical case several misfits occurred, for instance in the description of the pain and in the lab findings. However, according to a panel of four physicians the diagnosis of pancreatitis was still the most plausible, albeit in a more chronic and less vehement form than in the typical case.

Analysis

Think-aloud protocols

The analysis of the think-aloud protocols aimed at the identification of those parts of the protocols in which biomedical and clinical knowledge was applied in order to interpret

and diagnose the case. The identification of those parts was achieved in a step by step approach. The first step in the analysis of the think-aloud protocols was a rough segmentation based on pauses in the protocols. Next those segments containing more than one single 'basic conceptual operation' (e.g. generate a new hypothesis or verify an existing hypothesis, planning further information acquisition or identifying information need) were further subdivided, so each protocol segment may be assumed to represent one basic conceptual operation. Next, all segments pertaining to goal management and information need are excluded from the analysis as are segments pertaining to the perceived quality of the resulting problem representation (e.g. "I am not sure that what I am saying now is really right"). By doing so, a protocol-framework remained, consisting of segments in which a case finding was linked to an interpretation, one or more case findings were linked to a hypothesis (or vice versa) or in which two hypotheses were linked.

These remaining segments, represented as propositions consisting of (at least) two conceptual entities and a relation, were charted in semantic networks. In these networks, biomedical propositions were discriminated from non-biomedical propositions[2]. Criterion for this discrimination is the *object* of the proposition. Propositions concerning pathological principles, mechanisms or processes underlying the manifestations of a disease are classified as biomedical propositions. They are phrased in terms of entities such as viruses, bacteria, stones or carcinomas, in terms of tissue, organs, organ systems, or bodily functions. 'Irritation of peritoneum means diminished intestinal motility' is an example of such a proposition. By contrast, propositions concerning attributes of people, including their diseases, are labeled non-biomedical (Patel, Evans and Groen, 1989). These propositions are concerned with the ways in which a disease can manifest itself in a patient; the kind of complaints one would expect given a specific hypothesis; the nature and variability of the signs and symptoms and the ways in which the disease can be managed.

As the classification principle is based on the object of a proposition, often propositions from adjacent protocol fragments must be taken into account. The propositions were extracted and classified by two independent raters; whenever necessary, agreement was attained after discussion. The biomedical propositions were counted and this number was divided by the total number of extracted propositions. One audio recording (of subject #5-12, a fifth year student) contained so much noise that no transcription could be derived from it. Therefore, analyses of the think-aloud protocols were based on the data of 37 subjects.

Post-hoc explanations

The explanations of the underlying pathophysiological process were analyzed utilizing a method describe by Patel and Groen (1986). Patel and Groen segmented these texts into propositions consisting of two concepts and a relation. These propositions were represented as a semantic network and their number was counted.

Results

On-line knowledge application

The number of propositions extracted from the think-aloud protocols did not vary with an increasing level of expertise ($F(3,29)= 1.294$; $p= .2951$). However, the case variant diagnosed by the subjects strongly affected the number of knowledge application

---

[2] It should be noted that this classification biomedical - non-biomedical corresponds to the classification biomedical - clinical. In the way our classification system worked out non-biomedical was the default category. Hence, as far as the protocol analysis is concerned, the more technical term 'non-biomedical' is preferred.

propositions found in the think-aloud protocols ($F(1,29)= 8.821$; $p= .0059$). Figure 1 shows this effect. Apparently, diagnosing the atypical case required more knowledge application than the typical case.

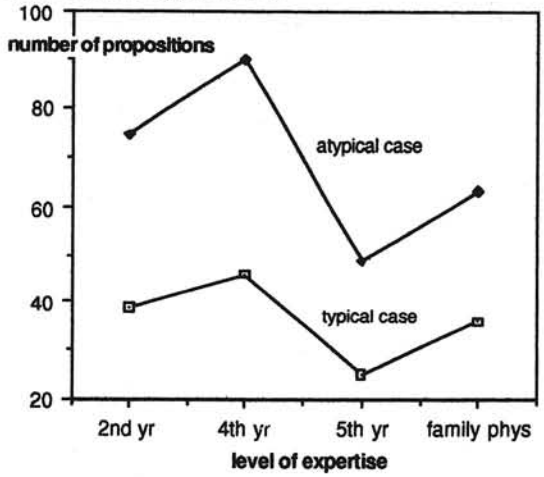


Figure 1. Number of knowledge application propositions extracted from the think-aloud protocols.

These knowledge application propositions were expressed at a varying number of case items. The number of items responded to varied with the subjects' levels of expertise ($F(3,29)= 2.856$, $p= .0542$) but did not vary with case type ($F(1,29)= .129$, $p= .7218$). Figure 2 shows that the fifth year students responded to the fewest number of items, indicating that these subjects were more selective than the other subjects.

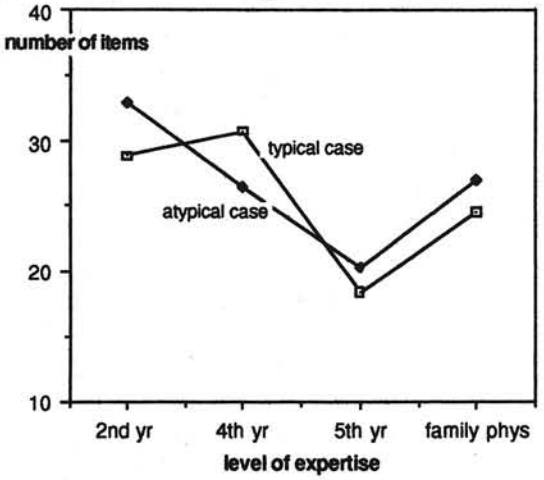


Figure 2. Number of case items responded to with knowledge application propositions

The share of biomedical knowledge in the total number of knowledge application propositions also varied with level of expertise ($F(3,29)= 5.196$, $p= .0054$), but not with

case type (F(1,29)= .712, p= .4056), nor an interaction of both factors was found (F(3,29)= .263, p= .8515). These effects are represented in Figure 3.

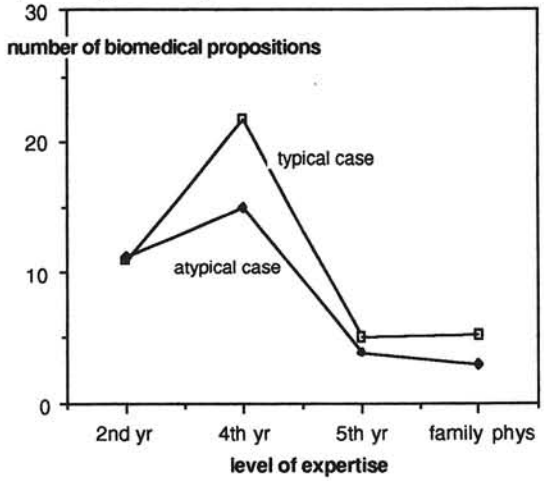


Figure 3. Number of biomedical propositions extracted from the think-aloud protocols.

In summary, subjects of different levels of expertise did not differ in the amount of knowledge applied in clinical reasoning. Notwithstanding that, level of expertise correlated with the number of case findings the subjects responded to with knowledge application propositions. Especially, fifth year students responded to a low number of case items, that is to say to less than half of them. Finally, the number of biomedical propositions also varied with level of expertise. Again this number was very low in the fifth year students, but the experts applied even less biomedical propositions. A peak was found in the fourth year students group. Practical experience seems the key to these differences between 2nd and 4th year students at one hand and 5th year students and experienced physicians at the other hand. So far these findings seem to confirm our hypothesis that the application of biomedical knowledge decreases with an increasing level of expertise, be it after an initial rise between the second and fourth year of study. However, this conclusion is complicated by another remarkable finding, regarding the difference in the number of knowledge application propositions applied while diagnosing the two different cases. Apparently, the atypical case required more cognitive effort. Notwithstanding that, the subjects did not apply more biomedical knowledge as was hypothesized.

Post-hoc knowledge application

The number of propositions in the post-hoc explanations was correlated with the subjects' level of expertise (F(3,30)= 4.168, p= .014). Figure 4 shows an almost monotonic increase with level of expertise. Increasing levels of expertise appear to be associated with a growth in the biomedical knowledge of pancreatitis and not with a decrease of the availability of this kind of knowledge as was hypothesized. Again, no differences related to case type were found (F(1,30)= .701, p= .4092). This finding is in sharp contrast with the finding that the on-line application of biomedical knowledge decreased after the fourth year level.

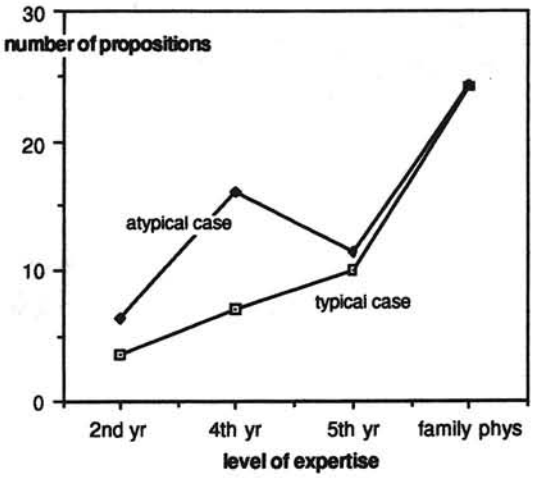


Figure 4. Number of propositions in the post-hoc provided pathophysiological explanations of the case.

Discussion

So far, some preliminary conclusions concerning our research questions can be drawn. First, our hypothesis that the application of biomedical knowledge decreases with increasing levels of expertise was confirmed, albeit after an initial rise between year two and four. The initial increase can be attributed to an increase in knowledge between year two and four. Second, investigation showed that this decrease is not caused by a decrease in availability of biomedical knowledge. Thus, we may conclude that the role of biomedical knowledge in expert clinical reasoning is virtually absent, while on the other hand this knowledge has not decayed. On the contrary, a steady growth of biomedical knowledge can be discerned.

Now the time had come to take the final, as yet unspecified step in our three step approach. This third step is needed in order to attain more insight in the organization of biomedical and clinical knowledge and in the mechanisms responsible for changes in the role and organization of clinical and biomedical knowledge.

Generally speaking two mechanisms can be hypothesized. The first possible explanation for this phenomenon is that expert biomedical knowledge has become inert in the course of clinical practice. The knowledge is still available in long term memory, as shown by the results of the post-hoc measurements, but simply is not used any more. Hence, experts would apply less biomedical knowledge in solving medical problems than intermediates. This would explain the apparent contradiction between the relative absence of biomedical concepts in the think-aloud protocols and their abundance in the post-hoc explanations.

The second possible explanation of the results is based on Anderson's theory of the development of cognitive skills (Anderson, 1983). According to Anderson (1983), students first try to solve problems in a specific domain applying elaborate (in this case biomedical) knowledge. Successful application of this elaborate knowledge, consisting of a chain of propositions, results in its compilation into a rule connecting problem features, to which this knowledge applies, and the outcome of the problem-solving process. In clinical reasoning, this compilation mechanism may result in the combination of sets of symptoms and their associated diagnosis.

In order to explore these two hypotheses, the overlap between applied and available knowledge was investigated. This amount of overlap was defined as the proportion of concepts in a subject's semantic network that were identical to any concept in the set of propositions derived from his or her think-aloud protocol. If biomedical knowledge becomes increasingly compiled with increasing expertise and is integrated in clinical knowledge, then a growing overlap of both kinds of knowledge is expected. If, however, biomedical knowledge becomes increasingly inert, no such increase in overlap is expected.

Overlap of think-aloud and post-hoc protocols

The proportion of concepts that appeared both in the post-hoc provided pathophysiological explanations and in the on-line applied knowledge varied with increasing levels of expertise ($F(3,29)= 14.977$, $p= .0001$). Figure 5 shows a monotonic increase with an increasing level of expertise. No effect of case typicality was found ($F(1,29)= 2.135$, $p= .1531$).

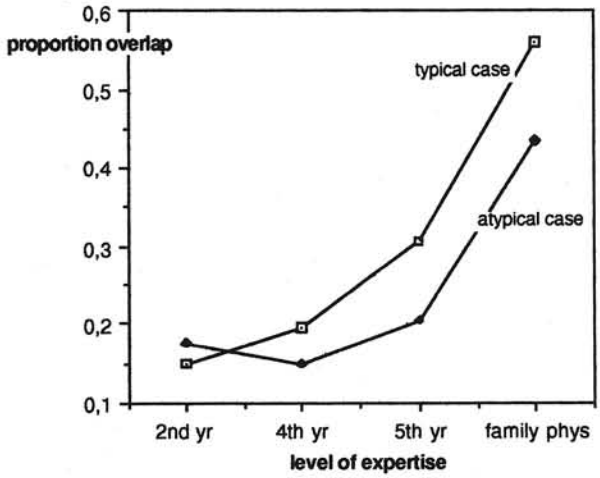


Figure 5. Proportion of common concepts in the think-aloud and post-hoc protocols

This finding contradicts the hypothesis that biomedical knowledge becomes increasingly inert and it is in agreement with the hypothesis of an increasing integration between biomedical and clinical knowledge. Hence, our analysis leads us to the conclusion that biomedical knowledge has not become rudimentary, nor inert, but instead becomes compiled and integrated in clinical knowledge.

Again the role of biomedical knowledge in clinical reasoning

Our results show that reasoning with clinical knowledge is preferred over biomedical knowledge in all levels of expertise. This observation does not disagree with our hypotheses. However, another observation does and that is the finding that our subjects applied *more clinical knowledge* in diagnosing an atypical case than in diagnosing a typical case. This finding was not expected, as biomedical knowledge was hypothesized to be needed for the explication of atypicalities in patient findings. In this paragraph we will try to explain this phenomenon.

For that reason we further investigated hypothesis generation and knowledge application in the think-aloud protocols (see table 1). These analyses showed no differences between the typical and atypical case in the moment the first hypothesis was brought for-

ward (F(1,28)= .163, p= .6891), although this moment tended to vary with level of expertise (F(3,28)= 2.64, p= .0689). Especially fifth year students tended to 'postpone' hypothesis generation. They needed about ten more items than the other subjects before a first hypothesis was brought forward.

TABLE 1
Hypothesis generation and diagnosis in the typical and atypical case.

	2nd year students		4th year students		5th year students		physicians	
	typical case	atypical case	typical case	atypical case	typical case	atypical case	typical case	atypical case
item# first hypothesis	11	10.5	9.5	12.5	19.25	20	8	9.6
item# pancreatitis first mentioned	25.8	–*	21	48	23	34	8	19.2
diagnosis **	.2	0	.5	1.75	1.5	2	1.6	1.4

* No 2nd year student mentioned the hypothesis of pancreatitis in the think-aloud protocols in the atypical case.

** Subjects were asked to give a differential diagnosis. If pancreatitis was mentioned as a first possibility 2 points were given, if pancreatitis was not mentioned at all no points were given, otherwise 1 point was given.

There were, however, strong differences related to case typicality in the moment the correct hypothesis (typical or atypical) pancreatitis was first mentioned (F(1,28)= 13.169), p= .0011). When the subjects tried to diagnose the atypical case, there was a delay of 15 items on the average, before the hypothesis 'pancreatitis' appeared. In the typical case, all physicians considered pancreatitis as one of the possible diseases that might cause the patient's complaint. This first hypothesis set was brought forward when the complaint was presented (item# 8). The content of this set of first hypotheses was highly influenced by case typicality and it took the physicians about ten items more on the average to come up with the hypothesis 'pancreatitis' in the atypical case. This discrepancy was even bigger in the student groups. For instance, the fourth year students typically furthered their first hypotheses after the 11th item had been presented. That is after the complaint and two additional items. The hypothesis 'pancreatitis' was furthered eleven items later in the typical case, but in the atypical case this hypothesis was only brought forward after the (atypical) lab findings (in the last item) had been presented. These lab findings seem to have changed their hypotheses set completely as is suggested by the final diagnosis. Two of the four fourth year students reported pancreatitis as their final diagnosis, the other two students reported it as a good second possibility. The fifth year students were even more convinced by the lab findings. All of them reported pancreatitis as a first diagnostic possibility. Remarkably, these students concluded more often to the diagnosis 'pancreatitis' when the atypical variant had been presented than in the typical case. The physicians on the other hand found pancreatitis a less likely diagnosis in the atypical case.

These results indicate that the atypical case requires much more information before the right hypothesis is generated and before the diagnosis is arrived at. Furthermore, they suggest that the students' mental representation and the associated hypothesis sets of the atypical case are less stable than in case of the typical variant. Apparently, biomedical knowledge is not used to interpret and order this "*unstructured*" mass of case information. Instead, clinical knowledge seems to be preferred for information ordering and in-

terpreting, while biomedical knowledge seems to be applied for a justification or explanation after the interpretation had been made.

An example of this way of reasoning is found in the think-aloud protocol of subject #4-15. After hearing the lab findings he concludes:

> "Serum amylase (32U) . increased .. that may indicate er a amylase is er . both er, let me think adrenaline amylase .. as ... hey wait a minute oh . wait that it just pops up .. the word pancreatitis .. er .. you don't have that that .. is specific for .. disease of the pancreas .. oh yes, sure alcohol .. the fact that er .. that pancreatitis is associated with alcohol consumption .. er yes high alcohol consumption .. that yes .. how was it exactly .. [some utterances about forgetting, having to study the subject again and not having thought of this hypothesis earlier] .. glucose 6.0 mmol/l. yes makes the pancreas more suspect .. if of course .. inflammation in the pancreas and er .. islets of Langerhans produce less insulin then . then of course a higher level of glucose remains [etc.]".

This example shows that first an item is clinically interpreted, while afterward a justification for this interpretation is construed. Most remarkably, this line of reasoning is set up to incorporate a finding that fits with the hypothesis generated. No such explanations are made in order to incorporate findings that do not really fit with the favorite hypothesis. This latter function for biomedical knowledge was however postulated. We must, however, keep in mind that in this experiment especially fourth year students applied biomedical knowledge. Nevertheless, the present findings raise the suspicion that theories that medical experts revert to biomedical knowledge when they have to diagnose a difficult case must at least be adjusted, if not completely reformed. As yet, however, the experimental results are not available to decide between these two options. An important prerequisite for this is to investigate medical experts solving difficult problems and applying biomedical knowledge in their own domain of expertise.

Conclusion

The presented experiment replicated the finding that (after an initial rise) the application of biomedical knowledge in clinical reasoning decreases with increasing levels of expertise. This decrease did not result from decay of biomedical knowledge. On the contrary, biomedical knowledge of the subject pancreatitis apparently increased with increasing levels of expertise. Furthermore, the analyses showed that biomedical knowledge had not become inert with increasing expertise. Finally, it was suggested that biomedical knowledge compiles and becomes increasingly integrated in the clinical knowledge base, resulting in a virtual absence of overt application of biomedical knowledge in the experts' think-aloud protocols.

Our theory on the role of biomedical knowledge in clinical reasoning was, however, complicated by two other findings. Biomedical knowledge was thought to be applied in order to accommodate deviating findings in the prevailing diagnostic hypothesis. The data did not support this assumption: Diagnosing the atypical case appeared to require more knowledge application propositions than the typical case, but, contrary to what was expected, an *equal* number of biomedical propositions was found. Differences in knowledge application resulted from an increase in the amount of clinical knowledge applied ($F(1,29)= 15.465$, $p= .0005$), while on top of that applied biomedical knowledge was used to explain why a matching instead of a deviating finding fitted with that hypothesis. Before any conclusions can be drawn from this result more specific research is needed.

References

1. Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

2. Boshuizen, H. P. A., Schmidt, H. G., & Coughlin, L. D. (1988). On the application of basic-science knowledge in clinical reasoning; implications for structural differences in knowledge between experts and novices. *Proceedings of the 10th annual conference of the Cognitive Science Society*. Montreal, Canada. Hillsdale, NJ: Erlbaum.
3. Boshuizen, H. P. A., Schmidt, H. G., & Coughlin, L. D. (1987). On-line representation of a clinical case and the development of expertise. Paper presented at AERA-conference Washington, D.C.
4. Feltovich, P. J., & Barrows, H. S. (1984). Issues of generality in medical problem solving. In H. G. Schmidt, & M. L. De Volder (Eds.), *Tutorials in problem-based learning; A new direction in teaching the health professions*. (pp. 128-142). Assen: Van Gorcum.
5. Kuipers, B. (1985). Expert causal reasoning and explanation. Paper presented at the Annual Conference of the American Educational Research Association. Chicago, IL. Chicago, IL.
6. Kuipers, B. J., & Kassirer, J. P. (1984). Causal reasoning in medicine; analysis of a protocol. *Cognitive Science, 8*, 363-385.
7. Lesgold, A. M. (1984). Acquiring Expertise. In J. R. Anderson, & S. M. Kosslyn (Eds.), *Tutorials in learning and memory; essays in honor of Gordon Bower*. San Francisco: Freeman & Comp.
8. Lesgold, A., Rubinson, H., Feltovich, P. J., Glaser, R., & Klopfer, D. (1988). Expertise in a complex skill: diagnosing X-ray pictures. In M. T. H. Chi, R. Glaser, & M. Farr (Eds.), *The nature of expertise*. Hillsdale, NJ: Erlbaum.
9. Patel, V. L., Evans, D. A., & Groen, G. J. (1989). Biomedical knowledge and clinical reasoning. In D. A. Evans, & V. L. Patel (Eds.), *Cognitive science in medicine; Biomedical modeling*. (pp. 53-112). Cambridge, Massachusetts: The MIT press.
10. Patel, V. L., & Groen, G. J. (1986). Knowledge-based solution strategies in medical reasoning. *Cognitive Science, 10*, 91-110.
11. Schmidt H. G. , Boshuizen H. P. A. , & Hobus P. P. M. (1988). Transitory stages in the development of medical expertise: the "intermediate effect" in clinical case representation studies. In: *Proceedings of the 10th annual conference of the Cognitive Science Society*. Montreal, Canada. Hillsdale, NJ: Erlbaum.

KENNISACQUISITIE VOOR EEN MEDISCH EXPERTSYSTEEM; THEORIE EN PRAKTIJK[1]

W. Krijgsman, J.H.C. Reiber, P.Fioretti, E.Backer[2], G.A. van der Ent[3], E.v.Royen[4]

Laboratorium voor Klinische en Experimentele Beeldverwerking, Thoraxcentrum, Erasmus Universiteit, Rotterdam.
[2]Vakgroep Informatietheorie, Fac. der Elektrotechniek, Technische Universiteit Delft.
[3]Stichting Sazinon, Meppel.
[4]Academisch Medisch Centrum, Afd. Nucleaire Geneeskunde, Amsterdam.

samenvatting
Formele kennisacquisitiemethoden zijn uitgebreid beschreven in de literatuur; helaas is er niet zoveel bekend over de problemen, die men in de praktijk tegenkomt en evenmin over manieren om deze problemen op te lossen. In dit artikel worden kennisacquisitie-ervaringen besproken, die zijn opgedaan bij de ontwikkeling van een expertsysteem t.b.v. de nucleaire cardiologie, te weten voor de analyse van Tl-201 scintigrammen. De ervaringen worden geschetst tegen de achtergrond van een formeel kennisacquisitiemodel, waarbij een aantal problemen worden belicht en waarvoor ook oplossingen worden aangedragen.

1. INLEIDING

Thallium-201 (Tl-201) scintigrafie is een nucleair-geneeskundige beeldvormingstechniek, die routinematig wordt toegepast voor de niet-invasieve bepaling van de regionale doorbloeding van de hartspier direct na maximale lichamelijke inspanning (gewoonlijk op een fietsergometer) en vier uur later, in de rustsituatie. Tl-201 is een radiofarmacon dat intraveneus wordt toegediend op het moment van maximale inspanning en zich via de bloedbaan verspreidt over het lichaam. Het Tl-201 wordt opgenomen door spierweefsel, dus ook het hartspier-weefsel, afhankelijk van de lokale doorbloeding en het metabolisme. Een vaste of roterende gamma camera wordt gericht op het hart om zodaning de door het hart en omliggend spierweefsel uitgezonden gammastraling te kunnen registreren. Deze informatie wordt vervolgens aangeboden aan een nucleair geneeskundig computersysteem, waarin overeenkomstige beelden kunnen worden gevormd in matrices van 64x64 of 128x128 beeldpunten. In de planaire Thallium scintigrafie worden achtereenvolgens opnamen vanuit drie richtingen gemaakt. In de tomografische Thallium scintigrafie roteert de camera over 180 of 360 graden om het hart, waarbij om de 6 graden een opname wordt gemaakt. Op basis van deze 30 of 60 aanzichten kan dan een drie-dimensionale verdeling van het Tl-201 in de hartspier worden gereconstrueerd. Ten behoeve van een gestandaardiseerde kwalitatieve en kwantitatieve beoordeling van de Tl-201 distributie worden vervolgens dwarsdoorsneden loodrecht op en parallel aan de lange as van het hart berekend. Dit resulteert in een totaal van twaalf plakken: 6 korte as doorsneden, 3 vertikale- en 3 horizontale lange as doorsneden.
De resulterende beelden tonen de gecumuleerde Tl-201 opname in de hartspier, hetgeen representatief is voor de regionale bloeddoorstroming. Door vergelijking van de overeenkomstige doorsneden na inspanning en bij rust kan de cardioloog beoordelen of er sprake is van normaal functionerend spierweefsel (normale doorbloeding, zowel bij inspanning als bij rust), een geïnfarceerd gebied (sterk verminderde doorbloeding zowel bij inspanning als bij rust), dan wel ischemie (een gebied met verminderde doorbloeding bij inspanning, maar een normale doorbloeding bij rust). Op basis van o.a. deze informatie wordt dan de verdere behandeling van de patient bepaald.

probleembeschrijving

Interpretatie van de beelden is moeilijk en vereist een lange leerperiode. Zo wordt aanspraak gedaan op de vaardigheden van de cardioloog of nucleair geneeskundige om zich een 3-dimensionaal beeld van het hart te vormen, en dan te bepalen of de gevonden defecten in de diverse beelden al-dan-niet consistent zijn. Herbij is het belangrijk dat de beoordelaar kleine verschillen in de helderheden in overeenkomstige beelden nauwkeurig kan onderscheiden. In dit proces moet hij rekening houden met de technische aspecten van de beeldvorming, alsmede met het ziektebeeld van de patiënt, om artefactuele defecten van echte defecten te kunnen onderscheiden.

---

1 Dit onderzoek is gesteund door de NWO, het gebiedsbureau voor de medische wetenschappen (subsidienr. 900-537-028)

De visuele interpretatie van de Thallium tomogrammen blijkt gepaard te gaan met grote intra- en interobserver variaties. Teneinde de defecten op een objectieve en meer reproduceerbare wijze te kunnen beoordelen, is een softwarepakket ontwikkeld voor de kwantitatieve analyse van Tl-201 tomogrammen [6]. Toch blijkt dat de cardioloog, naast de interpretatie van de kwantitatieve gegevens, veel belang hecht aan de visuele beoordeling van de beelden. Beoordelingsvariaties zijn kleiner geworden maar blijven bestaan.

introductie AI technieken
In ons streven naar een nog meer objectieve en reproduceerbare beoordeling is gekozen voor de ontwikkeling van een expertsysteem om zo ook de interpretatie-aspecten mee te kunnen nemen. De bedoeling is niet om de cardioloog te vervangen door een analyseprogramma, doch veeleer om hem te voorzien van extra gereedschap om meer consistente en reproduceerbare diagnoses te verkrijgen. Of, met andere woorden, het systeem dient de beoordelaar te begeleiden in de interpretatie van de beelden en van de kwantitatieve gegevens en fungeert als criticus in de totale analyse van de gegevens. Daarnaast kan het expert systeem ook gebruikt worden als leersysteem in de opleiding van nucleair cardiologische beoordelaars.

organisatie artikel
In dit artikel worden kennisacquisitie aspecten besproken tegen de achtergrond van het ESATS projekt (Expert Systeem voor de Analyse van Thallium-201 Scintigrammen). Eerst wordt een projektbeschrijving geschetst. In het kader van dit artikel zal alleen aandacht worden geschonken aan het kennisacquisitie aspect van dit projekt. Dan volgt de formulering van een formeel kennisverwervingsmodel. Hierna wordt de aanpak in het ESATS projekt besproken met een vergelijking naar het formele model. Er worden dan een aantal problemen beschreven alsook de gevolgde aanpak om tot oplossingen te komen. Tenslotte worden een aantal resultaten besproken.

2. PROJEKTBESCHRIJVING

Het projekt behelst de produktie van een expertsysteem voor de analyse van tomografische Thallium-201 scintigrammen, genaamd ESATS. Dit omvat de volgende stappen:

- o De definitie en produktie van een expertsysteem shell, welke geschikt is voor klinisch gebruik
- o De definitie en produktie van een kennisbestand
- o De definitie en produktie van additionele software voor :
  - het uitvoeren van externe routines, die worden geactiveerd vanuit het kennisbestand.
  - het lezen van patiëntgegevens uit een databank
  - het schrijven van de analyseresultaten naar een databank
  - het lezen van kwantitatieve beeldgegevens
  - het maken van een diagnoserapport
- o Statistisch onderzoek met een patiënten databank om onbekende relaties vast te stellen.
- o Evaluatie van het produkt.

Aan het projekt is de randvoorwaarde verbonden dat het expertsysteem operationeel moet worden op een standaard PC zodat verspreiding van het produkt in dit opzicht geen probleem mag zijn.

3. KENNISACQUISITIE: METHODOLOGIE

Het doel van kennisverwerving is kennis over een kennisdomein te modelleren. Het is hiervoor noodzakelijk om de structuur van het kennisdomein te ontdekken. Deze bestaat uit verzamelingen elementen, hun onderlinge relaties, eigenschappen en rand(voor)waarden. De structuur moet dan worden afgebeeld in een model. Het model wordt geëvalueerd, afgebeeld in een kennisrepresentatie en vervolgens geïmplementeerd en getest.

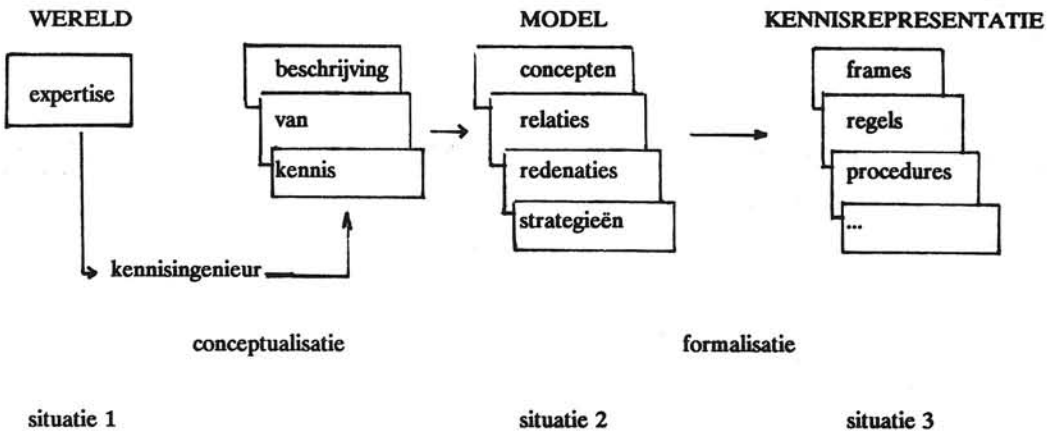


Fig. 1. Taak van de kennisingenieur: het verkrijgen en afbeelden van de expertise in een model en vervolgens in een kennisrepresentatie.

De gehele kennisacquisitie is een moeizaam proces, waarin de kennisingenieur maar al te vaak de bottleneck is (Fig. 1). Er zijn methodologieën ontwikkeld (voortgekomen uit de systeemontwikkeling) om dit proces zo optimaal mogelijk te laten verlopen [1,2,3,4,5]. Hiervan wordt een abstract model geschetst, dat de conceptualisatie van het kennisdomein toont, weergegeven in Fig. 1 door de overgang van situatie 1 naar situatie 2. Situatie 1 is de beginsituatie waarin de kennisingenieur nog niet de expertise heeft beschreven. Situatie 2 is een toestand, waarbij er een model is van de kennis in de vorm van een beschrijving, maar waarin deze beschrijving nog niet is gevat in een kennisrepresentatie. Het kennisacquisitieproces bestaat uit een drietal hoofdactiviteiten:

- o Het verkrijgen van kennis (van experts, enz.)
- o Het verwerken van delen kennis in een model
- o Het analyseren cq. evalueren van het model

Deze activiteiten vormen de hoofdbestanddelen van het model (Fig. 2) en worden hieronder nader beschreven.

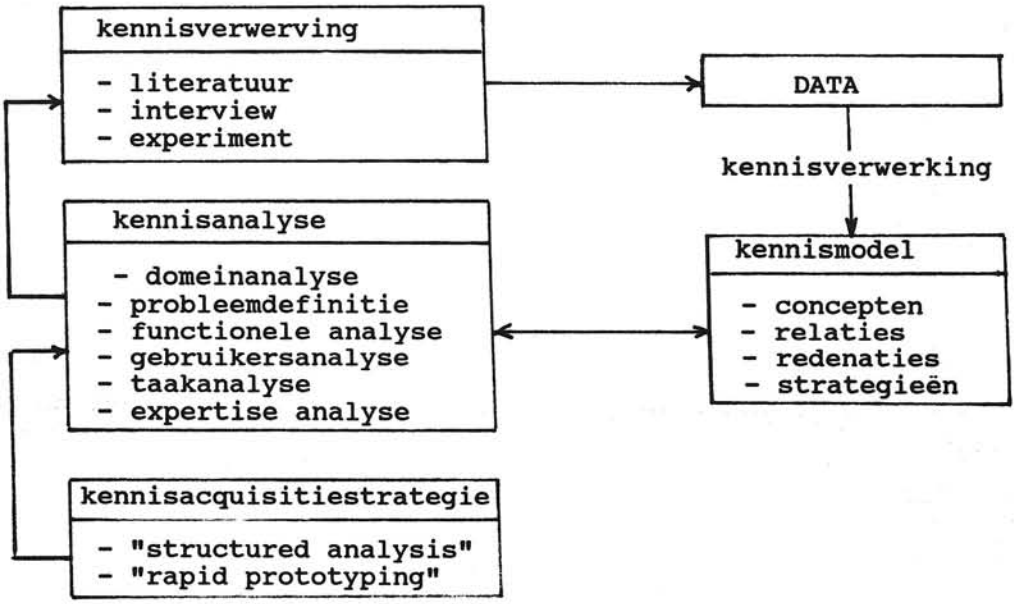


Fig.2 Schematische weergave van de kennisacquisitiemethodologie

o Het verkrijgen van kennis omtrent het kennisdomein

Vaak genoemde kennisverwervingstechnieken zijn:

- Het bestuderen van de relevante literatuur. Dit voorziet in algemene kennis over het domein, maar geeft zelden aan hoe problemen in de dagelijkse praktijk worden opgelost.
- Het houden van interviews. Interviews kunnen in principe wel deze informatie verschaffen, maar leveren vaak situatie-afhankelijke kennis, en geen algemene kennis. Er zijn veel verschillende vormen van interviews, die elk bepaalde typen kennis opleveren; een uitvoerige beschrijving vindt men o.a. in [1].

Een hier toegevoegde techniek is:

- Het uitvoeren van experimenten. Interviews leveren geen inzicht over praktijkhandelingen en -verrichtingen op grotere hoeveelheden situaties (in ons geval patiënten). Zorgvuldig ingerichte experimenten kunnen wel deze inzichten verschaffen. De inbreng van de expert is hier absoluut noodzakelijk om te zorgen, dat men inderdaad de juiste metingen uitvoert. In een experimentele opzet kunnen ook meerdere experts al-dan-niet tegelijk deelnemen.

o Het verwerken van delen kennis in een model

Het model wordt vastgelegd in modelgrootheden. Deze zijn in hiërarchische ordening van basis naar top:

- Concepten. Definieer de concepten van het domein, geordend in groepen en evt. geordend in een hiërarchie.
- Relaties. Bepaal hoe concepten met elkaar in verband staan. De relaties kunnen zowel statisch van aard zijn (algemeen geldig voor het domein) als afhankelijk van de specifieke situatie zijn.
- Redenaties. Bepaal hoe de expert relaties gebruikt om verbanden tussen concepten te leggen.
- Strategieën. Bepaal hoe de expert problemen aanpakt en oplost.

o Het analyseren cq. evalueren van het model

De aandachtsgebieden in de analyse zullen verschuiven in de loop van het kennisacquisitieproces. De analysestappen zijn:

- Domeinanalyse
- Probleemdefinitie
- Functionele analyse
- Gebruikersanalyse
- Taakanalyse
- Expertise analyse

In de analyse wordt het model steeds verder verbeterd en verfijnd. Het doel is hier na te gaan waar het model nog tekort schiet en incompleet is. Het model wordt hiervoor vanuit verschillende gezichtspunten bestudeerd. Op deze punten wordt dan gezocht naar aanvullende of aangepaste gegevens door herhaald kennisverwervingstechnieken toe te passen. Zijn gegevens bekend geworden, dan worden deze in het model ingepast. Dit gebeurt door voor elk additioneel of nieuw gegeven aandacht te schenken aan de compleetheid van en de consistentie met de overige modelgrootheden.
Er is dus sprake van een iteratief proces: in de evaluatie wordt herhaald aandacht geschonken aan de kennisverwerking, in de kennisverwerking worden de kennisverwervingstechnieken herhaald toegepast.

Het zij overigens opgemerkt, dat het model een afbeelding is van een stukje realiteit en als zodanig een kennisrepresentatie vormt. Dit is echter niet wat men in het algemeen onder deze term verstaat. Tot nu toe is 'alleen nog maar' de expertise vertaald van situatie 1 naar situatie 2; in figuur 1, de zg. conceptualisatie. Dan volgt de overgang van situatie 2 naar situatie 3, de formalisatie. Pas in situatie 3 is er werkelijk sprake van een kennisrepresentatie. Er bestaat, voor zover bekend, geen methodiek voor het kiezen van een kennisrepresentatie. Vaak zal een representatie worden opgelegd door de ontwikkelomgeving of door het expertsysteem. Indien er keuzemogelijkheden zijn, wordt dit aan het gezonde verstand van de kennisingenieur overgelaten. Ook in dit artikel zal geen methodiek worden besproken.

In de volgende paragrafen wordt de ESATS projektaanpak gevolgd en beschreven. De lezer wordt hierbij uitgenodigd de theorie interpretatie van de auteur met de praktijk te vergelijken en zijn conclusies te trekken...

4. KENNISACQUISITIE: PRAKTIJK

De eerste stap betrof een literatuurstudie op het gebied van de planaire Thallium-201 scintigrafie. De informatie hieruit werd vervat in een kort verslag. De inhoud bestond uit: de gevolgde procedures in het Thallium onderzoek, de beeldacquisitieprocedure, de reeks van beeldbewerkingsoperaties die op de originele beelden werden toegepast en resulteerden in de beelden die de cardioloog beoordeelt, een overzicht van parameters aan de hand waarvan de uitkomsten van de studies worden beschreven, welke patiënten zo'n onderzoek ondergaan, en welke klinische informatie dit onderzoek oplevert. Ook is een medische vocabulaire gemaakt.

Daarna werd een literatuurstudie uitgevoerd, volledig gericht op de Tl-201 tomografie. Tevens werd nu met grotere regelmaat overlegd met de lokale expert. Deze gesprekken waren zeer informeel en vonden in het algemeen plaats op zijn werkplek, terwijl hij bezig was met het diagnostiseren van patiëntenstudies.

Conclusies van deze eerste literatuur- en interviewfase waren:

- o De interpretatie van de Thallium beelden geschiedt volgens een vaste reeks van welomschreven stappen.
- o Scintigramanalyse is een eenvoudige procedure volgens de cardioloog.
- o De expert vindt het evenwel erg moeilijk duidelijk te maken wat hij precies 'doet' (d.w.z. het mentale proces) gedurende de diagnose. De gesprekken leidden tot te vage uitdrukkingen.
- o De expert is bereid zijn bezigheden te becommentariëren, zodat de kennisingenieur zelf de interpretatie van de beelden kan begrijpen.
- o Beeldinterpretatie lijkt de voornaamste bron van beoordelingsvariaties.
- o Op basis van de pre-test likelihood, die volgt uit de patientgegevens, en de Thallium test uitslagen kan een post-test likelihood worden bepaald. De Thallium test wordt dan gebruikt als een additionele meting die onafhankelijk wordt verricht van de overige testen. Dit blijkt niet overeen te stemmen met de praktijk omdat over het algemeen de cardioloog namelijk zijn patiënten kent. Hierdoor wordt de beeldinterpretatie beïnvloed door de patiëntgegevens zodat deze dubbel worden verdisconteerd en er dus een bias optreedt.
- o In veel artikelen wordt de waarde van de Thallium test als een indicator voor coronairlijden genoemd (vernauwingen in de kransslagaderen). Dit wordt gezien als een belangrijke rol van de test. Niettemin is de expert van mening, dat meer gedetailleerde uitspraken over de plaats en de ernst van coronaire obstructies in de dagelijkse praktijk moeilijk te doen zijn. Hij is ook niet te verleiden tot dergelijke uitspraken, en op de beoordelingsformulieren worden dergelijke voorspellingen niet vastgelegd.
- o Over het algemeen geldt, dat wanneer de beelden abnormaal zijn, d.w.z. wanneer de beelden hartdefecten tonen, de patiënt vervolgens een invasieve behandeling zal ondergaan.

Samenvattend: Het kennismodel bevat een reeks van geordende concepten. De analyse stappen zijn gedefinieerd, en de taken liggen vast, zij het dat hierover nog onzekerheden bestaan. Het is nog onduidelijk welke redeneringen de expert volgt in de interpretatie van de beelden. De redeneerstappen zijn ook nog grotendeels ongedefinieerd. Verder bestaat nog grote onduidelijkheid over de afleiding van coronairlijden uit de Thallium test uitslagen. Alles wat de meerwaarde van het expertsysteem t.o.v. het kwantitatief analysepakket moet bepalen, ligt nog open.

tweede fase

Om de sensitiviteit, specificiteit en diagnostische nauwkeurigheid van het expertsysteem te kunnen bepalen in vergelijking met de conventionele interpretatie met of zonder kwantitatieve gegevens, is een klinische evaluatie noodzakelijk. Daarom werd besloten een databank op te zetten, waarin routinematig alle Thallium-201 tomografische studies uitgevoerd op het Thoraxcentrum worden opgeslagen. Er volgde een lange periode van overleg met de experts over welke variabelen (lees: patiëntgegevens en Thallium test gegevens) wel en welke niet moesten worden opgeslagen. De discussies concentreerden zich vaak over specifieke gegevens die men wilde hebben en de benodigde inspanningen, om al die gegevens te verzamelen. Ook was van lang niet alle gegevens duidelijk of ze bij de klinische evaluatie van belang zouden kunnen zijn. Uiteindelijk, ruim een jaar na aanvang, is de databank in gebruik genomen. Een belangrijk neveneffect van deze exercitie is geweest, dat tegelijkertijd meer kennis is verkregen over het kennisdomein en dat het vocabulaire is verbeterd en uitgebreid.

Vanaf dat moment werden ook twee externe experts bij het projekt betrokken. De een uit een academische omgeving, de ander uit een perifeer ziekenhuis. Er werd besloten om regelmatig samen te komen, waarbij gemiddeld drie fysici en drie klinici aanwezig zouden zijn. Redenen hiervoor waren:

- o Sturing van het projekt
- o Brainstormen: het genereren en bespreken van ideeën over hoe kennis expliciet te maken.
- o Vaststelling van de definitieve functionaliteit van het expertsysteem.

De vermenging van klinici en fysici was gekozen om zo vanuit de diverse invalshoeken discussies aan te moedigen. Dat is zeker gelukt en de verschillende achtergronden van de klinici hebben daar veel aan bijgedragen.
Van alle besprekingen en experimenten werden volledige transcripties (van tape) gemaakt. Deze bleven exclusief eigendom van de kennisingenieur. Voor de overige leden van de groep werden aparte verslagen gemaakt die themagewijs de besproken onderwerpen, zinvol geachte uitspraken, conclusies en beslissingen bevatten, zonder te refereren naar degene die de betreffende uitspraken had gedaan; een vorm van discretie.
In de eerste bespreking kwamen de volgende punten aan de orde:

- o Voorstel voor de Thallium scintigram interpretatiestappen in het expertsysteem,
- o Aansluiting van het expertsysteem bij de klinische praktijk,
- o Inter- en intra-observer variaties,
- o Beoordelingsproces: verschillen en overeenkomsten tussen experts, is er sprake van redeneren of het herkennen van situaties, en hoe wordt statistiek toegepast,
- o Voorstel om een experiment uit te voeren: probeer om via een experiment de grootte van beoordelingsvariaties alsook de oorzaken van de verschillen te achterhalen, zodat dit een tipje van de sluier oplicht voor wat betreft het redeneren.

Samenvattend: Er zijn een aantal onzekerheden over de functionaliteit van ESATS verwijderd. En het kennismodel is verder ingevuld. Redeneerkennis en strategiekennis is nog niet aanwezig.

eerste experiment
Doel van het eerste experiment was om minstens een maal "expertise in action" te zien in een wat grotere opzet. Op grond hiervan zou worden vastgesteld in welke mate er sprake is van observervariaties. Dit kan informatie verschaffen over de oorzaken van de variaties en in hoeverre zij belangrijk zijn, hetgeen een bijdrage aan de beschrijving van expertise vormt. Het experiment werd uitgevoerd in drie afzonderlijke sessies, een sessie per expert. Elke sessie werd vastgelegd op tape. Het patiëntenmateriaal bestond uit tien Thallium-201 tomografische studies, zonder additionele gegevens, die volledig willekeurig waren gekozen door de kennisingenieur. De experts werd gevraagd de studies te beoordelen en hierbij hardop te denken.
Er werden standaard beoordelingsformulieren gebruikt, afkomstig uit een der instituten. De beoordeling vond plaats in twee stappen. Ten eerste werden de defecten visueel gescoord naar lokatie in segmenten en naar de ernst op een vijfpuntsschaal. Ten tweede werden dan de scores samengevat in een defectscore naar soort en naar ernst per gebied. Er werd een transcriptie gemaakt van de tape, en de uitslagen werden verzameld.

bespreking resultaten eerste experiment
Discussie van de resultaten van dit experiment was het onderwerp van de volgende bespreking. Het zij hier opgemerkt, dat de namen van de experts omwille van de discretie niet in de resultaten zijn genoemd; zij werden aangeduid met expert A, B en C. Dit is tot en met het laatste experiment volgehouden, ondanks het feit dat de experts dit niet nodig vonden (en hun identiteit ook onthulden).
De experts hadden commentaar op de experimenten. Zo zouden ze niet representatief zijn voor de dagelijkse gang van zaken, omdat er geen patiëntgegevens beschikbaar waren. Bij de routinematige beoordeling wordt immers rekening gehouden met de anamnese van de patiënt. Bij een eventueel volgend experiment diende een welgedefinieerde patientenpopulatie te worden gebruikt.
Er werden grote verschillen gevonden tussen de beoordelingen van de experts. Door alle tapes af te luisteren werd duidelijk welke redenaties zij volgden. Een aantal verschillen werden hierdoor verklaarbaar. Zo werden soms beelden niet meegenomen in de beoordeling, omdat ze te slecht van kwaliteit waren of omdat ze niet bruikbare doorsneden van het hart weergaven. Ook werden als afwijkend aangemerkte gebieden door de experts verschillend beoordeeld; soms werden afwijkingen a) beoordeeld als defect, b) toegekend aan de morfologie van het hart, c) verwaarloosd, d) samengenomen met andere defecten, e) beoordeeld als artefact, of e) soms niet opgemerkt.
De volgende stap was nu ten eerste te achterhalen of er een patroon kon worden vastgesteld voor de situaties, waarin deze verschillende interpretaties worden toegepast. En ten tweede te achterhalen of de experts het in dergelijke situaties eens konden worden over een interpretatie.
Het mag duidelijk zijn, dat hiermee een goed aanknopingspunt was gevonden om expertise te ontdekken.

Op grond van dit experiment werd duidelijk dat het expertsysteem de specifieke beeldkenmerken die aanleiding gaven tot de verschillende behandelingen van de afwijkingen, zou moeten extraheren uit de beelden. Voorbeelden hiervan zijn de richting van de hartassen, de grootte van de caviteit van het linker hartkamer, de morfologie van de afwijkingen en de morfologie van de hartspier in de overeenkomstige inspannings- en rustbeelden en tussen de doorsnedes onderling. Er zou gezocht moeten worden naar maten, die deze beeldkenmerken adequaat beschrijven. Bovendien zou voor elke maat een afbeelding gedefiniëerd moeten worden tussen de beeldkenmerkbeschrijving door de expert en de beeldkenmerkbeschrijving verkregen door kwantifikatie.

vervolgexperimenten

Er werden samen met de klinici nog twee experimenten gedefiniëerd. De bedoeling van de experimenten was om:

- o duidelijkheid te verkrijgen over de informatie, die de experts halen uit de kwantitatieve gegevens, welke geleverd worden door het kwantitatieve Thallium analysepakket,
- o een beeld te krijgen van de modificerende invloed van de patiëntgegevens op de uitslagen,
- o de betekenis van de verschillen in de beoordeling te bepalen,
- o de redeneerpaden achter die verschillende beoordelingen te achterhalen, en
- o te onderzoeken of de verschillende beoordelingen en redeneerpaden te verenigen zijn in een consensus.

Dezelfde drie experts werden ook betrokken bij deze twee vervolgexperimenten. Het materiaal bestond uit tien nieuwe patiëntenstudies, nu een geselecteerde groep patiënten. De experts werden gevraagd de studies te beoordelen, en hierbij hardop te denken. De beoordeling vond plaats in drie stappen. Eerst werd gevraagd de kwantitatieve data te interpreteren, vervolgens de beelden te analyseren en die mee te nemen in de diagnose, en tot slot ook nog de patiëntgegevens in de diagnose te betrekken. In het laatste experiment werden de experts gevraagd op basis van hun beoordelingen tot een consensusbeoordeling te komen.

bespreking resultaten vervolgexperimenten

Er was weer commentaar op de keuze van de patiëntenmateriaal. Terecht, want de patiënten bleken afkomstig uit een populatie die werd geanalyseerd volgens een bepaald protocol, en niet op aanvraag van een cardioloog. In een aantal situaties vond men, dat de kwantitatieve analyse onredelijk grote defecten liet zien, en daarop wilde men dus niet blindvaren. De additionele patiëntgegevens hadden een beperkte invloed; slechts in een klein aantal gevallen werd een diagnose gewijzigd. Er was verwacht, dat meer convergentie zou optreden in de diagnoses, naarmate meer gegevens beschikbaar kwamen. Welke rol de additionele gegevens moeten spelen in het expertsysteem is dus nog onduidelijk.

De verschillen in de beoordelingen zijn betekenisvol. Van iedere expert kon een 'beoordelingsgedrag' worden vastgesteld. De betekenis van de verschillen is door de experts zelf bepaald in het consensusexperiment. Er werd in alle gevallen consensus bereikt over het redeneerpad en de beoordelingen. In de consensusbeoordelingen was het mogelijk aan te geven welke beeldstructuren doorslaggevend waren voor de consensus.

Aan de hand van de expertbeoordelingen (met de consensusbeoordelingen als referentie) zijn de observer variaties bepaald (Fig. 3), alsook de sensitiviteit, specificiteit en nauwkeurigheid van de beoordelingen (Fig.4).

observer			CO	A	B
A					
	μ	s	-8.1		
		r	-7.3		
	σ	s	4.5		
		r	7.4		
	r	s	0.9		
		r	0.5		
B					
	μ	s	0.4	8.5	
		r	4.5	11.8	
	σ	s	11.1	10.3	
		r	10.1	10.8	
	r	s	0.7	0.8	
		r	0.5	0.3	
C					
	μ	s	-4.2	3.9	-4.6
		r	2.6	9.9	-1.9
	σ	s	10.1	10.2	14.4
		r	7.8	9.4	14.0
	r	s	0.8	0.7	0.5
		r	0.8	0.7	0.3

Fig. 3. gemiddelde μ, spreiding σ en correlatie r van de defect scores voor inspanning (s) en rust (r) van de drie beoordelaars vergeleken met elkaar en met de consensus beoordeling. De defect scores zijn gesommeerd over de visuele korte as defecten.
detectie van defecten

observer:	sensitiviteit	specificiteit	nauwkeurigheid
A	0.3-0.9	0.8-1.0	0.8-0.9
B	0.2-0.7	0.8-0.9	0.8-0.9
C	0.4-1.0	0.9-1.0	0.9-1.0

Fig. 4. Sensitiviteit, specificiteit en nauwkeurigheid van de detectie van een defect voor de drie beoordelaars t.o.v. de consensus beoordeling.

Fig. 3 geeft een indruk van het scoringsgedrag van elk van de experts. Zo scoort bijv. beoordelaar A gemiddeld duidelijk minder ernstig dan de overige beoordelaars. Uit Fig. 4 blijkt dat er, afhankelijk van het type defect,welke worden afgeleid uit de scores, vrij grote variaties bestaan voor de sensitiviteit in de detectie van een defect. Over het algemeen zijn de waarden van de specificiteit veel groter dan van de sensitiviteit. Kennelijk zijn de beoordelaars niet snel geneigd een (kleine) afwijking als zodanig te detecteren. Dit geeft aan dat voor de sensitiviteit in de detectie van defecten nog een significante verbetering mogelijk is. Een nauwkeurige kwantificatie kan hieraan bijdragen.

Samenvattend kan worden gesteld, dat de experimenten de volgende gegevens hebben opgeleverd:
- o de grootte van beoordelingsvariaties
- o Verschillen in de beoordelingen komen hoofdzakelijk voort uit verschillen in de interpretatie van de grijswaardeniveaus in de beelden; de relatie tussen grijswaarden in de beelden en afwijkingen in de doorbloeding van de hartspier.
- o Er zijn een aantal situaties geïdentificeerd (gebaseerd op bepaalde combinaties van beeldkenmerken) waarin defecten op een van de standaard afwijkende wijze kunnen worden beoordeeld.
- o Er kan voor deze situaties een consensusbeoordeling en een consensusredenatie worden bepaald.

De experimenten hebben geen enkele informatie opgeleverd over
- o de relatie tussen de defectdiagnoses en coronairlijden
- o de preciese invloeden van patientgegevens op de beoordeling van een Thallium studie

De redeneerkennis werd vervat in eenvoudig leesbare "IF ... THEN ..." tekstregels, zodat het mogelijk was deze met de experts te bespreken. Bovendien gaf dit hen een idee waar de kennisacquisitie toe leidde. Hieronder zijn een aantal voorbeelden van deze regels gegeven.

If a defect is present in only one slice
Then the defect is an artifact
= = A defect must be present in at least two slices to obtain any significance. Defects which are visible in only one slice are usually ignored. [5,8] = =

If the first short axis slice shows an anterior defect and
the defect (almost) disappeared in the second slice
Then the defect probably is artefactual
= = This is a difficult situation. The defect can be an artifact due to incorrect slicing; a reconstruction error. But the defect can be real, and is then probably caused by an obstruction in the septal branch of the LAD. [5] = =

If anterior or posterior defects found in the vertical long axis slices cannot be found in the short axis slices
Then the vertical long axis defects are ignored
= = If vertical long axis defects are not compatible with short axis defects, then the vertical long axis defects are generally not trusted and are thus ignored. = =

If no dipirydamole is used during exercise testing and
rapid washout (reverse redistribution) perfusion defects are found in the inferior wall
Then this possibly is an artifact due to splenic activity
= = Attenuation due to splenic activity manifests itself by an increase of tracer concentration in the inferior wall (visceral activity increases in delayed images). This mimics extensive ischaemia of the septum, anterior and lateral walls. [11, pp443]

5. DISCUSSIE EN CONCLUSIE

Het volgen van formele kennisaquisitie methodieken betekent niet automatisch dat een compleet model wordt verkregen. Zoals in het bovenstaande is beschreven, kan het voorkomen dat de kennisacquisitie als het ware 'vastloopt'. Het volgen van standaard methodieken is dus niet voldoende. De reden hiervoor is gelegen in het feit, dat de literatuur wel vertelt, **HOE** 'je het moet doen', maar niet vertelt niet **WAT** 'je moet doen'. De standaard interviewtechnieken leiden niet altijd tot de gewenste resultaten. Daarom is in het abstracte kennisacquisitiemodel de activiteit "experimenteren" opgenomen. Het bleek in dit projekt zeer zinvol enkele experimenten uit te voeren. De keuze van experimenten valt nauwelijks te modelleren. Wel is het zo dat onderzoek naar juist de verschillen in beoordelingen belangrijke aanknopingspunten kan opleveren.
In dit projekt is geen gebruik gemaakt van kennisacquisitie gereedschappen die uitgaan van een bepaalde methodologie, zoals het in ontwikkeling zijnde systeem KADS [1,3] of MORE [4]. Gebruik van dit soort gereedschappen is zeker aan te bevelen. Men moet zich echter wel bedenken, dat indien men zich niet conformeert of niet wenst te conformeren aan de gehanteerde methodologie, het hulpmiddel zich tegen de gebruiker zal keren, en daarmee eerder nadelen dan voordelen biedt.
De experimenten hebben in ieder geval de reeds bestaande indruk bevestigd, dat de visuele interpretatie in eerste instantie een patroonherkenningsprobleem is. Inter- en intra-observer variaties worden o.a. veroorzaakt door verschillen in interpretatie van helderheidsvariaties in de beelden, versterkt door het feit, dat referentiewaarden in de beelden ontbreken. Om de lokalisatie, grootte en ernst van defekten op een objectieve en reproduceerbare wijze te kunnen beoordelen, is het nodig een nauwkeurige kwantificatie uit te voeren. Ten behoeve van deze kwantificatie moeten morfometrische parameters worden bestudeerd en ontwikkeld. Van belang is ook, dat het expertsysteem storende structuren, vormafwijkingen, etc. herkent en hiervoor corrigeert. Het expertsysteem kan bovendien de scintigrambeoordelaar helpen om een vaste volgorde van interpretatie te volgen, waardoor de reproduceerbaarheid en betrouwbaarheid hopelijk verbetert. Het expertsysteem moet ten slotte ook kunnen uitleggen, op basis van welke feiten en kennis de interpretatie is vastgesteld.

Referenties

[1] Breuker, JA, Techniques for knowledge elicitation and analysis, Report 1.5, Esprit project 12, Amsterdam, July 1984: 16-36.

[2] Kidd, A, Knowledge elicitation for expert systems: a practical handbook, New York, 1988.

[3] Schreiber, G, Breuker, J, Bredeweg, B, Modelling in KBS development, 2nd. Eur. Knowledge acquisition workshop EKAW'88, Bonn, June 1988.

[4] Kahn, G, Nowlan, S, McDermott, J, Strategies for knowledge acquisition, IEEE Trans. Pattern An. and Machine Intell., vol PAMI-7, no.5, 1985: 511-522.

[5] Guida, G , Tasso, C, Topics in expert system design -methodologies and tools, North-Holland, Amsterdam, 1989.

[6] Reijs, AEM, Reiber, JHC, Fioretti, PM, Thallium-201 tomography: developments towards quantitative analysis, In:Signal Processing III:theory and applications, IT Young et al. (eds.), EURASIP, 1986: 1401-1404.

Een kennisgebaseerd systeem voor de automatische benoeming van bloedvaten op angiografieën

L Maes, D Delaere, C Smets, P Suetens, F Van de Werf

Katholieke Universiteit Leuven
Interdisciplinaire onderzoekseenheid voor radiologische beeldverwerking
(ESAT-MI2 + radiologie)
Kardinaal Mercierlaan 94
B-3030 Heverlee (Belgium)

Afdeling Cardiologie, UZ Gasthuisberg
Herestraat 49
B-3000 Leuven (Belgium)

1. Abstract

In dit artikel bespreken we de interactie met ezperten voor de ontwikkeling van een automatisch kennisgebaseerd systeem voor de interpretatie van bloedvaten op angiografieën. Nadruk ligt vooral op het verwerven van de anatomische kennis en de implementatie ervan. In een eerste stap wordt deze gehaald uit boeken en anatomische atlassen. Beeldopnames en protocoleringssessies worden bijgewoond om inzicht te krijgen in de handelingen en interpreteringsstrategieën. Vervolgens wordt een eerste poging tot implementatie ondernomen. Er wordt gebruik gemaakt van een regelgebaseerde taal, namelijk OPS5. Deze eerste versie wordt beknopt uitgelegd aan de cardioloog en de resultaten getoond, zodat hij voor de informaticus begrijpelijke en implementeerbare kritiek kan leveren. De nodige kennis wordt dus verkregen via een tweerichtings-proces. Daarin wordt afwisselend het geheel van de reeds overgedragen kennis geïmplementeerd en het resultaat hiervan laat de cardioloog toe de kennisbank te vervolmaken.

2. Inleiding

We beschrijven onze ervaringen bij de ontwikkeling van een regelgebaseerd beeldverwerkingssysteem voor het aflijnen, benoemen en ruimtelijk voorstellen van de linker kransslagader.

In een eerste stap worden de bloedvaten afgelijnd, gebruik makend van kennis van de radiografische projectie van een bloedvat.

In een tweede stap worden de gevonden bloedvat-segmenten benoemd met hun anatomische naam. Dit is dus het implementeren van medische, anatomische expertkennis. Hierbij wordt gebruik gemaakt van een Constraint Satisfaction algoritme. De anatomische kennis wordt geformaliseerd als constraints op locale attributen als plaats, richting en lengte enerzijds en op relaties tussen bloedvatsegmenten als "links van", "verbonden met" anderzijds. De linker kransslagader wordt benoemd in de twee standaard projecties. Er wordt geopteerd voor een robuust systeem dat bij

slechte beeldkwaliteit of segmentatiefouten geen foute benoemingen maakt, maar eventueel aan specificiteit kan verliezen.

In een derde stap wordt de kennis uit twee projecties gecombineerd. Bloedvaten uit beide projecties worden gecorreleerd op basis van naam, lengte en dikte. Deze correlaties laten ons onder meer toe het ruimtelijke verloop van de bloedvaten te bepalen.

We zullen ons toespitsen op hoe de anatomische kennis voor het benoemen van de bloedvaten met hun correcte anatomische naam werd vergaard en geïmplementeerd.

3. Anatomische kennis

Om zelf kennis over de anatomie van de linker kransslagader te verkrijgen maakten we in eerste instantie gebruik van anatomische atlassen. Dit gaf ons inzicht in de boomstructuur van de bloedvaten en in hun ruimtelijk verloop. Deze kennis bleek echter onvoldoende om zelf in klinische beelden alle bloedvaten juist te benoemen. In de atlassen wordt veel belang gehecht aan anatomische referentiepunten op het hart die op klinische angiografieën niet zichtbaar zijn. Voorbeelden hiervan zijn : " De hoofdstam splitst ter hoogte van de crux cordis in de LAD, die verder verloopt in de interventriculaire groeve, en de circumflex die in de atrioventriculaire groeve ligt". De atlassen geven bovendien om didactische redenen een geïdealiseerd beeld van de coronairboom. De belangrijke anatomische varianten worden vaak vereenvoudigd weergegeven. De studie van deze werken leverde voor ons wel belangrijke inzichten op, maar deze waren vaak moeilijk te gebruiken in klinische beelden, of moeilijk te implementeren. De beschreven drie-dimensionele boomstructuur gaat verloren door de radiografische projectie en door segmentatiefouten.

In tegenstelling tot de anatomische atlassen, die de nadruk leggen op het ruimtelijk verloop en de onderlinge samenhang van de kransslagaders, hebben wetenschappelijke werken specifiek over coronarografie een andere aanpak. De verschillende takken worden afzonderlijk beschreven. De nadruk ligt hierbij op eigenschappen van de 2-dimensionele projecties van deze vaten. De septaaltak wordt in RAO projectie beschreven als een weinig mobiele, dunne, rechte tak die ongeveer loodrecht op de LAD staat. Vele van deze beschrijvingen zijn vrij eenvoudig te implementeren aan de hand van een constraint satisfaction algoritme. Ook in deze werken werd vaak een geïdealiseerd beeld gegeven van de coronairboom.

Het bijwonen van opname- en protocoleringssessies maakte het ons mogelijk om de opgedane kennis aan de realiteit te toetsen. Ons inzicht in de ruimtelijke boomstructuur uit de anatomische atlassen, hielp ons de 2-dimensionele eigenschappen uit de coronarografie-tekstboeken zinvol toe te passen. We konden zo een beperkte ervaring opdoen, doch voldoende om een implementatie-strategie te ontwerpen.

4. Implementatie-strategieën

Voor de interpretatie van klinische coronarografieën maken we zelf vooral gebruik van beschrijvingen van 2-dimensionele projecties van losse takken. In een volgend stadium trachten we onze interpretaties te verfijnen aan de hand van onze kennis van de ruimtelijke boomstructuur. De meeste van deze beschrijvingen zijn eenvoudig als beperkingen of constraints te formuleren. We gebruikten dan ook een constraint satisfaction algoritme, waarbij elk gedetecteerd bloedvatsegment een verzameling van mogelijke interpretaties heeft. Aanvankelijk bevat deze verzameling alle mogelijke interpretaties, om na toepassing van de beperkingen slechts één juist element te bevatten.

4.1. Unaire constraints.

Deze beperkingen baseren zich op het feit dat wanneer we één enkel segment beschouwen, zonder rekening te houden met wat zich er rond bevindt, we toch al een vrij belangrijke kennis over dit segment hebben. Verscheidene interpretaties kunnen geschrapt worden. Voor elk segment hebben we reeds de volgende interessante kennis: RAO- of LAO-projectie, localisatie in het beeld, oriëntatie, lengte, dikte, intensiteit. Vele van deze constraints werden bijna letterlijk uit de tekstboeken over coronarografie overgenomen.

Enkele voorbeelden:

- RAO localisatie: een segment links-onder kan niet LAD zijn.
- RAO oriëntatie: een horizontaal segment kan niet circumflex zijn.
- lengte: een segment langer dan 2 cm kan niet de hoofdstam zijn.
- dikte/intensiteit: een segment dikker dan 5 mm kan geen septaaltak zijn.

Uiteraard zijn combinaties ook mogelijk en interessant:
bv. een segment dat in RAO rechts-onder ligt en verticaal is, kan geen septaaltak zijn.

Een voorbeeld van een dergelijke regel voor de LAO-projectie in OPS5 is:

```
(P SKIP_MAIN_LCA_1
  [(SEGMENT_LCA ^BEGINPOINT <BEGIN>
                ^ENDPOINT <END>
                ^MAIN_LCA  T)          <SEG>]
  [IT_IS_SO ^THAT (SEGMENT_LIES_IN <BEGIN> <END>
                      1 1 0 0 0 0 1 1
                      1 1 0 0 0 0 1 1
                      1 1 1 0 0 1 1 1
                      1 1 1 1 1 1 1 1
                      1 1 1 1 1 1 1 1
                      1 1 1 1 1 1 1 1
                      1 1 1 1 1 1 1 1
                      1 1 1 1 1 1 1 1 )]
  -->
  [MODIFY <SEG> ^MAIN_LCA  F] )

verklaring: - neem een segment dat nog de hoofstam als mogelijke
              interpretatie heeft.( ^MAIN_LCA  T(rue) ).
            - kijk of het in de opgegeven zone ligt.
            - indien dit zo is, schrap dan die interpretatie.
```

Deze unaire constraints zijn een eerste en belangrijke stap naar de correcte benoeming van de vaten. Zij vormen de basis waarop de relaties van de binaire constraints verder bouwen. Na toepassing van deze unaire constraints hebben we al een vaag idee over hoe de coronairboom eruit ziet. We hebben onze anatomische kennis uitgebreid: naast de algemene anatomische kennis over de linker coronair, hebben we nu ook een belangrijke kennis van de specifieke coronairboom waar we mee werken. De binaire constraints tesamen met de nu opgedane kennis van deze specifieke kransslagader, zullen de benoemingen verder specifiëren.

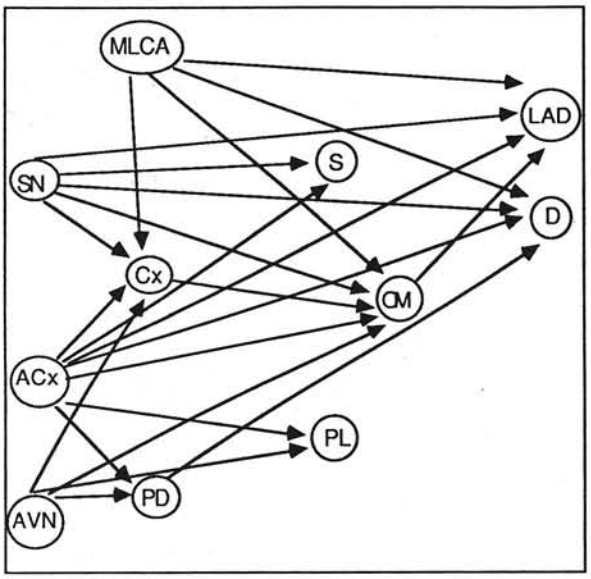


Fig. 1 : De relatie "ligt links van"

4.2. Binaire constraints

Bij de unaire constraints werd slechts één enkel segment tegelijk beschouwd. Onafhankelijk van hoe het verdere beeld eruit ziet, werden conclusies getrokken voor dit segment. Bij de binaire constraints zullen we relaties tussen twee segmenten beschouwen. Voorbeelden van deze relaties zijn: segmenten A en B maken deel uit van hetzelfde bloedvat; segment A ligt links van segment B; segment A takt af van segment B;....

Uit onze ruimtelijke boomstructuur en projectie-kennis weten we heel wat over relatieve localisaties van vaten. In RAO-projectie bijvoorbeeld, kunnen links van de circumflex slechts de hoofdstam, de atriale circumflex en/of de kleinere knooptakken liggen. Een volledig uitgewerkte relatie "ligt links van" wordt afgebeeld in figuur 1. Wanneer we dus een tak hebben met interpretatie "circumflex", kunnen we een ganse reeks interpretaties schrappen bij de segmenten die er links van liggen.

De relatie "takt af van" is diegene die op het eerste gezicht het meest voor de hand ligt. We weten immers dat de coronairen een boom vormen. Zoals reeds hoger vermeld gaat het in de projectie eerder om een willekeurige grafe. De relatie valt dan uiteen in beperkingen als "kan links/rechts verbonden zijn met" en " kan onder/boven verbonden zijn met".

5. Kennisuitbreiding en -verfijning

De hierboven beschreven aanpak stelde ons in staat een systeem te bouwen gebaseerd op eigen kennis en ervaring. Het is in staat om in relatief eenvoudige klinische coronarografieën de belangrijke takken juist te benoemen.

Door te opteren voor data-structuren waarmee op een hoog niveau over bloedvaten kan geredeneerd worden, is het systeem eenvoudig aanpasbaar en uitbreidbaar. Dit laat ook toe om de globale werking vlot aan de cardioloog uit te leggen. Zonder in technische details te treden is het voor hem mogelijk om zich vertrouwd te maken met de manier van redeneren zoals die geïmplementeerd is. Expert en programmeur kunnen nu in een voor elkaar verstaanbare taal over onvolkomenheden en verbeteringen overleggen.

Een iteratief proces werd nu gestart. De resultaten van de reeds geïmplementeerde regels worden voorgesteld aan de cardioloog. Deze bekritiseert deze resultaten. In overleg met de programmeur worden regels aangeduid die te streng zijn en dus fouten induceren en nieuwe regels voorgesteld om de specificiteit van het systeem te verhogen. Deze nieuwe informatie kan dan door de programmeur ingebracht worden.

Door dit proces te herhalen op een groeiende set beelden, wordt meer en meer expertkennis en ervaring in het systeem gebracht.

6. Besluit

In dit artikel stelden we voor hoe we expert-kennis verkregen voor de ontwikkeling van een automatisch systeem voor het interpreteren van coronarografieën. Studie van tekstboeken en het actief bijwonen van protocolerings-sessies maakten het ons mogelijk om een implementatie-strategie te kiezen. De nadruk lag hierbij op het ontwikkelen van data- en programmeerstructuren die het mogelijk maken op een hoog niveau te redeneren over bloedvaten. Hierdoor kon de ingebrachte kennis eenvoudig uitgebreid of verbeterd worden. De expert kon zich snel vertrouwd maken met de gekozen strategie en er ontstond een vruchtbare dialoog tussen expert en programmeur. Een iteratief proces werd gestart waarbij de reeds bekomen resultaten van het systeem door de cardioloog beoordeeld en verbeterd werden. Dit gaf aanleiding tot het steeds weer implementeren van nieuwe expertkennis en -ervaring.

7. Referenties

[1] GG Gensini, "Coronary arteriography" Futura Publishing Company inc., Mount Kisco, New York, 1975.

[2] Sobotta, "Atlas of Human Anatomy" Vol. 2, H Ferner & J Staubesand, Urban and Schwarzenberg, Munich-Vienna-Baltimore, 1982.

[3] F.N. Netter, " The Ciba collection of medical illustrations: The Heart", Case-Hoyt Corporation, Rochester, NY, 1978.

[4] Smets C., Suetens P., Van de Werf F., "A Knowledge-Based System for the Labeling of the Coronary Arteries." Proc. of the SPIE, Newport Beach, Feb 1989, to appear.

[5] S. Tsuji, H. Nakano, " Knowledge-Based Identification of Artery Branches in Cine-Angiograms - An Image Understanding System which Utilizes Produstion-Type Knowledge -" in Computer Science and Technologies, T. Kitagawa, ed., North-Holland, 311-321 (1982).

[6] A. K. Mackworth, "Constraint Satisfaction" Enceclopedia of artificial intelligence, ed. SC Shapiro, John Witley and Sons, 1987, p205-211.

[7] K. Barth, R. Koch, P. Marhoff, " Automated three-dimensional recognition of the coronary tree with clinical DSA image pairs. 14th International Conferencew On Computers In Cardiology ", Leuven 13-16 september 1987

[8] L. Brownston, R. Farrell, E. Kant, N. Martin, " Programming Expert Systems in OPS5. An introduction to rule-based programming ", Addison Wesley Publishing Company 1985.

[9] J.Y. Catros, D. Mischler, " An artificial approach for medical picture analysis", Pattern Recognition Letters 8, 123-130, 1988.

[10] J.L. Elion, S.E. Nissen, " A Knopwledge-based Image Procerssing System for the Interpretation of Coronary Arteriograms ", Proc. SPIE, Medical Imaging I, Vol 767, 428-432, 1987.

[11] M. Fischler, O Firschein, " Intelligence, the eye, the brain and the computer ", Addison Wesley Company, 1987.

De toepasbaarheid van technieken voor automatisch leren in medische domeinen: een case study

W. Post en M.W. van Someren

Vakgroep Sociaal Wetenschappelijke Informatica
Vakgroep Cardiologie
Universiteit van Amsterdam
Herengracht 196
1016 BS Amsterdam

1 Inleiding

Het is welhaast clichematig, maar daarom niet minder waar, te zeggen dat kennisacquisitie ten behoeve van kennissystemen buitengewoon lastig is. De hiervoor vaak gebruikte 'bottle neck'-metafoor geldt eens te meer voor de ontwikkeling van systemen in medische domeinen. Deze domeinen omvatten een groot aantal specialismen en het is moeilijk om de kennis die voor een bepaalde klasse van problemen nodig is, af te bakenen. Verder is redeneren met onzekerheid vaak een belangrijk element in medische systemen maar vooralsnog kan dit niet eenvoudig geïntegreerd worden met andere wijzen van redeneren. Twee kennisacquisitieproblemen waarnaar nog weinig onderzoek is verricht, zijn het *onderhoud* en de *aanpasbaarheid* van medische kennissystemen. Er komt regelmatig nieuwe informatie beschikbaar uit geneeskundig onderzoek die ingebouwd zou moeten worden in het kennisbestand en als een systeem in een andere context gebruikt moet gaan worden zijn eveneens aanpassingen nodig. Dit vergt extra kennisacquisitie. Tot slot

*Het hier beschreven onderzoek is mede gefinancierd door ESPRIT project P 2576 ACKnowledge en door de Nederlandse Hartstichting. De auteurs danken Michael Sramek en Ruud Koster voor hun medewerking.

zijn de eisen die worden gesteld aan medische kennissystemen vaak stringent. Foutieve antwoorden van het systeem kunnen fatale gevolgen kunnen hebben en verder blijft de gebruiker van het systeem in de meeste gevallen verantwoordelijk voor de uiteindelijke beslissing. In verband daarmee moet een systeem in veel gevallen zijn antwoorden kunnen uitleggen.

Er zijn drie benaderingen van het kennisacquisitieprobleem die elk een eigen oplossing voor het kennisacquisitie probleem inhouden.

De eerste benadering staat bekend als **prototyping** (zie [1]). Het is een bottom-up aanpak, die zich kenmerkt door een oppervlakkige analyse van het domein en een vroegtijdige start van de bouw van het systeem. Er zal snel resultaat geboekt worden op een onderdeel van het domein. De struktuur van de kennis en van het redeneerproces worden vaak meer bepaald door de implementatieomgeving dan door de struktuur die de kennis bij de expert heeft. Hierdoor worden vaak allerlei ad hoc beslissingen genomen en stuit men soms later in de bouw op problemen die alleen door aanpassing van het al aangelegde fundament op te vangen zijn. Dit zal zijn uitwerking hebben op de delen van het systeem die op dit fundament zijn gebaseerd. Ingewikkelde herstelwerkzamenheden of zelfs sloop kan het gevolg zijn. Verder leent de kennisbank die zo wordt opgebouwd zich doogaans slecht voor het geven van uitleg, omdat de vorm van de kennis niet goed aansluit bij de manier waarop gebruikers over het domein denken.

De tweede benadering is **gestruktureerde kennisacquisitie.** Het is een top-down aanpak, waarin eerst een conceptueel model wordt opgesteld van de expertise dat zich in een later stadium eenvoudig laat vertalen naar bouwstenen van de implementatie. Het kenmerkt zich door een gedegen analyse van het domein en het probleemoplosgedrag van de expert alvorens met de bouw te beginnen. De KADS methodologie ([1]) is hiervan een schoolvoorbeeld. De kennis die op deze manier verzameld wordt leent zich beter voor toepassing van principes uit de software engineering bij het bouwen van het programma dan prototyping.

De derde benadering berust op het gebruik van **gevalsbeschrijvingen.** Van deze benadering bestaat diverse varianten. Een bekende vorm is de statistische benadering (Bayesiaanse statistiek en multivariate technieken). Het gebruik van kwantitatieve in tegenstelling tot kwalitatieve data staat hier centraal. (zie bijv. [2], blz 263 en [9]).

Een andere variant van deze benadering berust op het zoeken naar symbolische kennis in een domein. Het gaat hier primair om begrijpelijke, symbolische relaties en niet zozeer om algebraische funkties. Deze variant gaat onder de naam *automatisch leren* en komt voort uit de Kunstmatige Intelli-

gentie. Deze laatste variant is het onderwerp van deze studie.

Door middel van een vergelijkend onderzoek gaan we na in hoeverre het gebruik van leertechnieken een levensvatbare methode vormt voor kennisacquisitie in medische domeinen. Dit alles vindt plaats in het kader van de ontwikkeling van een systeem voor de herkenning van acute hartziekten. Wij beperken ons hierbij tot technieken die automatisch leren van voorbeelden. Deze keuze wordt bepaald ten eerste door het feit dat de aard van de taak zich er goed voor leent (het betreft een klassificatie-taak), ten tweede door het feit dat het verzamelen van kennis in de vorm van voorbeelden in onze situatie eenvoudig uitvoerbaar is, en ten slotte doordat we te maken hebben met een domein waarvan de toepassingssituatie nogal aan veranderingen onderhevig is. Voor dit laatste kan een adaptief systeem uitkomst bieden.

2 Automatisch Leren

Technieken voor automatisch leren zijn grofweg in vier groepen onder te verdelen. De eerste groep omvat het automatisch leren door *inductie*, d.w.z. het afleiden van algemene wetmatigheden uit een verzameling feiten. De tweede groep betreft *operationalisatie* van kennis. Hierbij wordt effektief toepasbare kennis afgeleid uit theoretische kennis. De derde groep technieken leert door *instructie*. Hierbij is er sprake van een externe 'docent' die kennis ook letterlijk doceert, d.w.z de kennis gestruktureerd aanbiedt of helpt struktureren i.p.v de strukturering aan de leertechniek overlaat. In de vierde groep technieken staat leren door *analogie* centraal. Hier wordt getracht om oplossingen voor een bepaalde taak bruikbaar te maken voor nieuwe, min of meer vergelijkbare situaties.

Het idee om technieken voor automatisch leren toe te passen in medische domeinen is op zich niet nieuw. Er zijn bijvoorbeeld pogingen ondernomen op het gebied van lymphografie, oncologie, etc. ([6]). Er is nog maar weinig vergelijkend onderzoek gedaan naar de vraag welke techniek onder welke omstandigheden voor medische domeinen het meest geschikt is.

Het onderhavige onderzoek betreft het automatisch leren van voorbeelden. Er zijn twee technieken gebruikt die automatisch leren door inductie. De technieken, AQ (zie [6]) en ID3 ([7]), maken gebruik van voorbeelden en tegenvoorbeelden van een bepaalde klasse. Beide technieken leren algemene herkenningsregels voor het herkennen van klassen. AQ leert herkenningsregels, door telkens uitgaande van een voorbeeld van een klasse, generalisaties

van dat voorbeeld te vormen die niet in strijd zijn met de bekende negatieve voorbeelden (de gevallen die bij andere klassen horen). De generalisatie die volgens bepaalde criteria (parameters van de techniek) het beste is, wordt bewaard. Op deze manier worden alle positieve voorbeelden afgewerkt. De gevonden deel-geenralisaties worden gecombineerd tot een herkenningsregel. ID3 bouwt een beslisboom voor de bepaling van het klasse lidmaatschap. Dit gebeurt door telkens een attribuut te kiezen en de voorbeelden te verdelen naar de waarde die ze op het attribuut hebben. Door deze procedure telkens recursief toe te passen op de ontstane deelverzamelingen wordt een beslisboom gebouwd. Als alle voorbeelden in een bepaalde tak van de boom tot dezelfde klasse behoren, wordt deze klasse met de betreffende tak in de beslisboom geassocieerd. Beide technieken zijn speciaal voor dit onderzoek geïmplementeerd.

3 Wijze van evaluatie

Om een uitspraak te kunnen doen over de toepasbaarheid van technieken voor automatisch leren in medische domeinen zullen daarvoor eerst een aantal evaluatiecriteria voor de verworven kennis moeten worden opgesteld. Buchanan en Shortliffe ([2]) geven aan dit probleem aandacht en ook anderen hebben zich hiermee bezig gehouden (zie bv. Hollnagel ([5]), Fieschi en Joubert ([4])). In dit verband zijn er verschillende wijzen van vergelijking van belang. In het gunstigste geval is er een absoluut criterium voor handen, een gouden standaard, waaraan een oplossing van een systeem kan worden getoetst. In de medicijnen levert autopsie het uiterlijke, zij het nogal rigoreus verkregen, criterium. Er zijn echter vaak ook diagnostische testen met een voldoende hoge betrouwbaarheid om als absoluut criterium te kunnen worden beschouwd.

Ten tweede is de vergelijking met mensen die dezelfde taak uitvoeren van belang. Hiermee wordt aangegeven wat er in de praktijk gehaald wordt. Het geeft ook een aanknopingspunt voor welk niveau voor een machinale techniek haalbaar of acceptabel is. Een belangrijk probleem bij deze benadering is wel dat domein experts het onderling oneens kunnen zijn. In dat geval wordt gekeken of verschillen tussen voorspellingen van het systeem en voorspellingen door experts gemiddeld even groot zijn als verschillen tussen experts onderling.

Het acquisitieproces kan geëvalueerd worden door de resultaten van technieken te vergelijken. Dit wordt gedaan met behulp van kruisvalidatie: met

een deel van de data worden regels geleerd en het resultaat wordt op een ander deel getoetst. Behalve de prestatie (aantal goed herkende gevallen), kunnen de sensitiviteit en de specificiteit bepaald worden.

4 Het domein en de data

De context waarin de vergelijking plaats vindt, is de ontwikkeling van een systeem voor de automatische diagnostiek van pijnklachten op de borst, waarbij het hartinfarct, angina pectoris, functionele klachten en diverse ritmestoornissen de meest frekwente diagnoses zijn.

Het primaire doel van het systeem is het bieden van ondersteuning voor centralisten van ambulancediensten bij de beslissing om een ambulance uit te sturen of niet. De geboden haast bij dergelijke beslissingen is evident. Beslissingen zijn vanwege het feit dat er alleen verbale communicatie plaatsvindt alleen te nemen op basis van anamnestische informatie, zoals informatie over klachten (pijn, benauwdheid), symptomen en risicofactoren (geslacht, leeftijd e.d.) etc. Een diagnose is in tegenstelling tot de gangbare medische praktijk hier duidelijk niet gebaseerd op uitgebreid diagnostisch onderzoek, zoals analyse van het electrocardiogram (ECG) en bloedonderzoek.

Voor de verzameling voorbeelden aan de hand waarvan de inductietechniek de herkenningsregels voor ons te bouwen systeem moest afleiden was een bestand van gevalsbeschrijvingen van patiënten met pijnklachten op de borst speciaal voor dit project aangelegd. Hiertoe was een dialoogprogramma geschreven waarmee de cardiale anamnese kon worden afgenomen. De anamnese van een cardioloog stond hiervoor model en de dialoog werd gecomplementeerd met vragen naar ander informatie die volgens de literatuur over pijnklachten op de borst en eigen onderzoek mogelijk relevant zouden zijn. In totaal werden 45 kenmerken opgenomen. Sommige kenmerken hebben binaire waarden (*ja/nee* kenmerken als *pijn op de borst*), andere zijn nominaal (kenmerken als *hevigheid van de pijn* met waarden *zeer hevig, weinig hevig, niet hevig*) en weer andere zijn meervoudig van karakter, wat wil zeggen dat ze meerdere waarden tegelijk kunnen aannemen (het kenmerk *plaats van de pijn* met als waarden *links op de borst, midden op de borst, linkerarm, etc.*).

Met dit programma werd in de loop van een half jaar de cardiale anamnese afgenomen van zo'n 350 patiënten met pijnklachten op de borst die zich aanmeldden op de Eerste Harthulp (E.H.H.) van het Academisch Medisch Centrum bij de Universiteit van Amsterdam of aldaar op de Hartbewaking

werden verzorgd. Op de E.H.H. presenteren zich zo'n tien á twintig patiënten per dag. Hiervan wordt ongeveer 60 % na observatie naar huis gestuurd (de observatie duurt in principe niet langer dan 24 uur).

Het dialoogprogramma bleek mede vanwege de uitgekiende struktuur en doordat het volledig muisgestuurd is uitstekend dienst te doen. Administratieve handelingen waren nauwelijks meer nodig. Bovendien heeft het een belangrijk voordeel dat mogelijke ruis door onvermijdelijke typefouten tot een minimum beperkt kon blijven.

Aan elke patiëntbeschrijving is achteraf de ontslagdiagnose toegevoegd en door twee experts aan de hand van het gehele opname dossier nogmaals gecontroleerd. De betrouwbaarheid van de klassificaties verschilt enigzins per klasse. De diagnose hartinfarct kan objectief worden vastgesteld door middel van een bloedonderzoek. De betrouwbaarheid is zo goed als absoluut. Mocht een inspannings ECG positief hebben uitgewezen dan is ook ischaemie vastgesteld. Soms ligt dat anders. Zo wordt soms de diagnose *mogelijk angina pectoris* gesteld, waarmee de onzekerheid daarover impliciet wordt aangegeven. En functionele klachten wordt meestal per uitsluiting gegeven. In het algemeen kan worden gezegd dat hoe zieker de patiënt is hoe zekerder de diagnose kon worden gesteld, dus hoe betrouwbaarder het klasse label is.

In het totaal werden 360 patiënten geinterviewd waarvan 104 in de klasse *hartinfarct* ondergebracht werden, 59 in de klasse *instabiele angina pectoris* (waaronder de diagnose *dreigend hartinfarct* is vervat), 44 in de klasse *stabiele angina pectoris*, 70 in de klasse *functionele klachten*, 19 in de klasse *overige acuut-cardiale gevallen* (waaronder diagnoses als klepgebreken, aneurisma etc.), 36 in de klasse *supraventriculaire tachicardie* en 28 in de klasse *overige niet-acute gevallen* (waaronder diverse diagnoses als groep, nitrobaatcollaps e.d. resorteren).

In de hieronder gepresenteerde resultaten is een onderverdeling in urgente en niet-urgente gevallen gemaakt, met respectievelijk 182 en 178 voorbeelden. De acute gevallen waren de patiënten met *hartinfarct* of *instabiele angina pectoris* en de *overige acuut-cardiale gevallen*. De niet-acute gevallen zijn alle overige patiënten.

5 Experimenten en resultaten

Om de toepasbaarheid te beoordelen hebben we de absolute en relatieve waarde van de door AQ geproduceerde herkenningsregels en de door ID3 geproduceerde beslisboom onderzocht. De absolute waarde kan bepaald

worden door middel van kruisvalidatie, waarbij de herkenningsregels op een deel van de verzamelde voorbeelden werden geleerd en op een ander deel getest. De relatieve waarde met betrekking tot menselijke experts hebben we bepaald door twaalf experts ieder twaalf voorbeelden uit onze verzameling voor te leggen met de opdracht ze te klassificeren als zijnde urgent of niet. De relatieve waarde met betrekking tot de andere kennisacquisitie methoden hebben we bepaald door ook een gangbare statistische techniek, namelijk logistische regressie-analyse, toe te passen en door een variant van de prototyping-methode toe te passen op hetzelfde domein.

De logistische regressie-analyse werd als volgt uitgevoerd. Ten eerste werden op basis van chi-kwadraten de belangrijkste kenmerken geselecteerd. Vervolgens werd via een stapsgewijze analyse gezocht naar een optimaal regressie-model voor de helft van onze voorbeeldenverzameling. Uiteindelijk bleven een model met 10 over. Met dit model werden de voorbeelden uit de andere helft van de voorbeeldenset geklassificeerd.

De prototyping-variant ging als volgt. We lieten een expert eenvoudige vuistregels voor het domein opstellen. Deze vuistregels hadden de vorm van een conjunctie van beschrijvingskenmerken die een klasse lidmaatschap impliceert. De kenmerken waren exact dezelfde als die waarmee de verzamelde voorbeelden waren beschreven. Twee voorbeelden van deze regels zijn:

```
klachten = pijn &
bekend_met_pijn = ja &
hevigheid_pijn = zeer hevig &
erger_dan_anders = ja &
nitrobaat-effect = geen_effect
   ==> urgent

klachten = (pijn en hartkloppingen)
& bekend_met_pijn = nee
   ==> niet urgent
```

Deze regels werden vervolgens interaktief getest op onze verzameling voorbeelden. Deze test gaf per regel feedback over hoe correct de regels de verzameling voorbeelden klassificeerden (hoe vaak een negatief voorbeeld voor een positief voorbeeld werd aangezien) en hoe volledig de verzameling voorbeelden geklassificeerd kon worden (hoeveel voorbeelden niet geklassificeerd konden worden). Dit gaf informatie om de regelset uit te breiden c.q. in te krimpen of te verfijnen. Verfijning vond simpelweg plaats door kenmerken toe te voegen of weg te halen.

In tabel 1 zijn de prestaties van de verschillende methoden en de menselijke experts naast elkaar gezet.

Tabel 1
Overzichtstabel

	AQ	ID3	Prototyping	Statistiek	Experts
Acuut vs. Niet Acuut					
Correct in %	64	58	67	68	68
P-apriori	51	53	50	50	53
Sensitiviteit	.69	.64	.73	.85	.84
Specificiteit	.58	.52	.62	.50	.48

De prestaties van alle benaderingen komen ruim boven het niveau uit wat op grond van de apriori kans (P-apriori) wordt verwacht. De sensitiviteit is steeds hoger dan de specificiteit wat inhoud dat acute gevallen beter worden herkend dan niet acute. Bij de experts is dit veruit het duidelijks, wat wellicht betekent dat men in geval van twijfel het zekere voor het onzekere neemt. Het resultaat van de logistische regressie-analyse behoeft nog enige toelichting. Het regressie-model werkt met een instelbare drempelwaarde die het verband tussen de sensitiviteit en de specificiteit, zoals uitgedrukt in een ROC-curve, kan vastleggen.

Bij de prototyping-variant dient nog vermeld te worden dat bij het zoeken naar de beste regelset ook steeds een afweging plaatsvond, namelijk tussen een zo hoog mogelijke correctheid en een zo groot mogelijke compleetheid ten aanzien van de voorbeeldenset. Uiteindelijk (na verschillende sessies over enkele weken verspreid) werd een optimale regelset bereikt die voor 87 % van de *geklassificeerde* voorbeelden correct bleek maar slecht 46 % van alle voorbeelden *kon* klassificeren. In bovenstaande tabel zijn deze twee percentages gecombineerd door de 54 % niet geklassificeerde gevallen voor 50 % goed te rekenen (gelijk de apriori kans op correcte klassificatie).

6 Conclusies

Uit de resultaten blijkt dat technieken voor automatisch leren een interessant alternatief vormen voor kennisaquisitie in medische domeinen naast 'prototyping' en een statistische benadering. De prestaties zijn vergelijkbaar met die van de andere methoden en bovendien wordt het niveau van menselijke experts benaderd.

De absolute prestaties die met de verworven kennis bereikt worden, zijn niet bijzonder goed. Slechts zo'n 2/3 van de voorspellingen bleken correct. Voor de verklaring hiervan moeten we kijken naar het domein en de leervoorbeelden. Weliswaar zijn de voorbeelden zorgvuldig en op uniforme wijze verzameld, maar er zijn een aantal andere faktoren die zowel mensen als inductietechnieken voor problemen plaatsen.

Waarschijnlijk is het verband tussen de beschikbare kenmerken van de patiënt en de diagnose niet bijzonder sterk. De verzameling kenmerken is samengesteld op grond van literatuuronderzoek en uitvoerige gesprekken met experts, maar het is toch mogelijk dat er kenmerken ontbreken. In dit domein zit er naar alle waarschijnlijkheid een limiet aan de prestatie. Zoals eerder is genoemd zijn gegevens uit verdergaand medisch onderzoek, zoals het ECG en het bloedonderzoek, buiten beschouwing gelaten. Op basis van ECG en bloedonderzoek zijn betere diagnoses mogelijk. Bovendien geven de experts tijdens het beoordelen van de 12 aangeboden gevallen ook aan dat ze sommige informatie in de voorbeelden misten. Uitbreiding van de beschrijving lijkt dan ook gewenst, voor zover dit mogelijk is in het domein van telefonisch herkenning van hartklachten.

Als we de verschillen tussen de technieken bekijken, zien we dat die vrij klein zijn. Er is een verschil in prestatie tussen de twee inductietechnieken, AQ en ID3. We gaan hier niet verder op in. Prototyping en statistiek leveren hier vergelijkbare resultaten op. We merken op dat in de tabel met resultaten de sensitiviteit en specificiteit gegeven zijn. Deze zijn echter gemakkelijk te manipuleren bij zowel statistische als automatisch leer-technieken.

Bij de keuze voor een bepaalde benadering en daarbinnen voor een bepaalde techniek is niet alleen de kwaliteit van de antwoorden na een acquisitieproces van belang. Moet het systeem uitleg kunnen geven dan is men genoodzaakt om de gestruktureerde kennisacquisitie te plegen, waarbij een gedegen analyse van het domein moet worden uitgevoerd. Noch inductie, noch prototyping, noch statistiek vormen een voldoende basis voor een uitlegfaciliteit. Wordt deze eis niet gesteld dan genieten technieken voor automatisch leren mogelijk de voorkeur boven prototyping omdat dit minder van een expert vergt. Het gebruik van inductietechnieken is relatief goedkoop, de geproduceerde regels zijn inzichtelijker dan algebraïsche funkties en daardoor aanknopingspunten geven voor verdere kennisacquisitie direkt van een expert. Een voorwaarde is dat de voorbeelden al voorhanden zijn of eenvoudig verzameld kunnen worden. Wat dat betreft zal er een kosten-baten analyse aan een techniekkeuze vooraf moeten gaan.

Voordelen van automatisch leren en statistiek zijn, dat ze direkt gebruikt

kunnen worden voor verder onderhoud en voor aanpassing van het systeem aan een nieuwe, mogelijk afwijkende toepassingscontext. Men verzamelt nieuwe gevallen en die kunnen worden gebruikt voor een nieuwe leersessie of ze kunnen worden toegevoegd aan de oude, waarna opnieuw geleerd kan worden, al naar gelang de situatie.

Een interessante mogelijkheid is om verschillende benaderingen te combineren. In plaats van inductie pur sang, kan men inductie trachten te combineren met interaktieve elicitatie bij de expert of met het bouwen van een gestruktureerde kennisbank. Een mogelijkheid is om inductie te gebruiken voor het verfijnen van gestruktureerde kennis. Een systeem dat dit doet is INDE ([8]). Symbolische technieken zijn hierbij beter toepasbaar dan statistische. Een analyse van de mogelijkheden van combinaties van technieken is te vinden in [3].

Een andere mogelijkheid is dat geïnduceerde regels worden door een expert verfijnd en gestruktureerd in een elicitatiesituatie. Dit levert een variant op van de prototype benadering. In onze prototype situatie maakte de expert al gebruik van evaluatie van zijn regels op voorbeelden, maar door het grote aantal voorbeelden is dat lastig. Inductie zou dit proces kunnen ondersteunen.

In veel toepassingen is expliciete kennis nodig voor uitleg en voor onderhoud van het systeem. De symbolische regels die de inductietechnieken opleveren en het symbolische karakter van het inductieproces bieden daarvoor waarschijnlijk een goed uitgangspunt. Een voorbeeld hiervan is het gebruik van speciale, kansrijke generalisaties. Uit een klein experiment ([10]) bleek dat voorkennis over het domein in de vorm van betekenisvolle interpretaties (bv. "pleurale prikkeling", afleidbaar uit samenhang van de pijn op de borst met ademhaling en bewegingen van de romp) door het AQ algorithme kan worden gebruikt voor het vinden van even adequate, maar meer begrijpelijke generalisaties.

References

[1] J. A. Breuker and B. J. Wielinga. Model Driven Knowledge Acquisition. In P. Guida and G. Tasso, editors, *Topics in the Design of Expert Systems*, pages 265–296, Amsterdam, 1989. North Holland.

[2] B.G. Buchanan and E.H. Shortliffe. *Rulebased Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project.* Addison Wesley, Reading, Massachusetts, 1984.

[3] B. Wielinga et.al. Conceptualisation of a knowledge engineering workbench. Technical Report ACK-UvA-T1.4-DL-010-A, Universiteit van Amsterdam, Amsterdam, 1990.

[4] M. Fieschi and M. Joubert. Some reflexions on the evaluation of expert systems. *Methods of Information in Medicine*, 25:15–21, 1986.

[5] E. Hollnagel. Evaluation of expert systems. In P. Guida and G. Tasso, editors, *Topics in the Design of Expert Systems*, Amsterdam, 1989. North-Holland.

[6] R.S Michalski, I. Mozetic, J. Hong, and N. Larsson. The aq15 inductive learning system, an overview and experiments. Technical Report ISG 86-20, Dep. of Comp. Science, Un. of Illinois, Illinois, Urbana, 1986.

[7] J.R. Quinlan. Consistency and plausible reasoning. In *IJCAI-83*, pages 137–144, 1983.

[8] P.P. Terpstra and M.W. van Someren. Inde: leren met induktie en deduktie. In A.Th. Schreiber and M.W. van Someren, editors, *NAIC'88*, pages 185–196, Amsterdam, 1988. SWI.

[9] J.H. van Bemmel. Formalization of medical knowledge. *Methods of Information in Medicine*, 25:191–193, 1986.

[10] P. van der Velden. Experimenten met een lerend systeem. stageverslag, Universiteit van Amsterdam, Amsterdam, 1990.

MEDISCHE BESLISSINGSONDERSTEUNING: DE RELEVANTIE VAN ONTWERPBESLISSINGEN VOOR DE ACQUISITIE VAN MEDISCHE KENNIS.

R.B.M. Jaspers [1]
ITI-TNO, Delft

Inleiding

Lange tijd zijn medische beslissingsondersteunende systemen (mbos) gebaseerd geweest op impliciete kennis afkomstig van gegevens van het te ondersteunen medisch proces. Dit waren eenvoudige systemen voor databank analyse (Fries, 1972; Starmer et al, 1979), of classificatie systemen gebaseerd op het theorema van Bayes of op patroonherkenning (Warner et al, 1964; Gorry en Barnett, 1968). Daarnaast werden met behulp van systeemidentificatietechnieken systemen ontwikkeld voor applicaties in de geneeskunde waar statische modellen van het proces niet voldoen, bijvoorbeeld voor prognose of behandelplanning (Blom, 1975; Stassen et al, 1980). De mogelijkheid deze 'gegevensgebaseerde' systemen te ontwikkelen was afhankelijk van de beschikbaarheid van voldoende gegevens van het medisch proces. Zeker met het complexer worden van de mbos werd data-acquisitie een belangrijk probleem (Jaspers, 1990). Door de opkomst van AI-technieken die het realiseren van medische kennissystemen mogelijk maakten werd aan dit probleem voor medische beslissingsondersteuning een eind gemaakt. Medische kennis kon expliciet in het mbos worden gerepresenteerd zonder dat deze uit gegevens moest worden geëxtraheerd. Hiervoor in de plaats kwam echter een ander probleem, namelijk dat van de acquisitie van de benodigde kennis.

Acquisitie van medische kennis

In dit artikel wordt onder kennisacquisitie verstaan het proces van kennisvergaren uit tekstboeken en via elicitatie van experts, de analyse en interpretatie van deze kennis en het vormen van een conceptueel model van het medisch proces. Kennisacquisitie wordt door velen gezien als de bottleneck voor het realiseren van medische kennissystemen. Desalniettemin blijft dit probleem relatief onderbelicht. Neale (1988) stelt vast dat in de literatuur betreffende eerste generatie expert systemen met name het maken van een conceptueel model van het proces nauwelijks aan bod komt. Naar zijn waarneming heeft het er alle schijn van dat bij het realiseren van deze systemen deze fase nauwelijks heeft plaatsgevonden, maar dat vergaarde kennis rechtstreeks werd vertaald naar produktieregels. Dit wordt deels veroorzaakt door het feit dat het opstellen van een conceptueel model nauwelijks wordt ondersteund. Ondersteuning bij het

---

1) Het onderzoek waarover in dit artikel wordt gerapporteerd is door de auteur uitgevoerd bij de vakgroep Werktuigkundige Meet- en Regeltechniek van de TU-Delft met subsidie van het Delfts Universiteits Fonds en het Praeventiefonds.

ontwikkelen van kennissystemen richt zich voornamelijk op het implementatieniveau. In dit artikel zal worden getoond dat voor het realiseren van mbos het ontwerp van deze systemen moet worden ondersteund op een hoger abstraktieniveau dan het implementatieniveau. Het belang hiervan voor de kennisacquisitie zal worden besproken.

Medische kennissystemen

Medische kennissystemen zijn gebaseerd op expliciete kennis van het medisch proces. In vergelijking tot andere domeinen waarin kennissystemen worden toegepast zullen medische kennissystemen doorgaans gebruik moeten maken van een grote verscheidenheid aan typen kennis, ieder met zijn eigen specifieke representatievorm:
- Kennis betreffende anatomische strukturen.
- Causale kennis van fysiologische processen.
- Heuristische kennis van fysiologische processen waarvan de exacte werking niet bekend is.
- Causale associaties betreffende het verband tussen symptomen en oorzaken.
- Strategische kennis voor het efficient oplossen van problemen.

Dit wordt veroorzaakt door het feit dat medische kennis dikwijls onzeker is en incompleet. Dikwijls wordt een onderscheid gemaakt tussen wat wordt genoemd 'oppervlakkige' kennis, bestaande uit heuristieken of causale associaties die zijn gebaseerd op ervaring met het medisch proces (know how) en 'diepe' kennis, gebaseerd op inzicht in deze processen, bestaande uit beschrijvingen van de struktuur van processen en de fysische of fysiologische wetten waarmee het gedrag van deze strukturen kan worden beschreven (know why). Beide soorten kennis hebben specifieke eigenschappen die relevant zijn voor het ontwikkelen van kennissystemen en voor de kennisacquisitie.
Diepe kennis kan dikwijls worden gevonden in tekstboeken zonder de tussenkomst van experts. Dit leidt tot een kort, gestruktureerd ontwikkeltrajekt. Oppervlakkige kennis daarentegen dient deels te worden verkregen van experts. De eerste generatie kennissystemen was voor een belangrijk deel gebaseerd op heuristische kennis, waarbij symptomen op een hoog niveau worden geassocieerd met oorzaken. Deze 'heuristische klassifikatie' (Clancey, 1985) biedt een efficiënte specifieke redeneerstrategie voor diagnostische problemen. Het gebruik van empirische associaties heeft echter ook een aantal nadelen. Deze betreffen tekortkomingen in de uitlegfaciliteiten van het resulterende systeem en mogelijk onvoorspelbaar gedrag aan de rand van het domein. Een groot bezwaar wordt daarbij gevormd door het feit dat het nodig is een complete verzameling empirische associaties te vinden teneinde problemen aan de rand van het domein te voorkomen. Hierdoor zijn de ontwikkelkosten van systemen gebaseerd op causale associaties hoog. Dit heeft geleid tot interesse in systemen die gebaseerd zijn op diepe kennis van het domein. Echter, door de incompleetheid van medische kennis is diepe kennis omtrent een domein dikwijls niet volledig. In dat geval dient voor het ontwikkelen van kennissystemen ook van expert heuristieken gebruik gemaakt te worden. Gezocht dient dan te worden

naar een methode van kennisacquisitie, -organisatie en -representatie waarbij de genoemde nadelen zo veel mogelijk worden vermeden.

Voor een suksesvolle introduktie van medische kennissystemen dient in het algemeen aan een aantal eisen te worden voldaan. Naast algemene software engineering eisen van onderhoudbaarheid, performance en kort ontwikkeltrajekt dienen medische kennissystemen tegemoet te komen aan specifieke gebruikerseisen die betrekking hebben op de kwaliteit van het advies en inzichtelijkheid van het systeem in verband met mogelijke uitlegfaciliteiten (Van Daalen, 1988). Dit vereist een expliciete keuze van een groot aantal facetten die liggen op een hoger abstraktieniveau dan dat van de implementatie:
- Het niveau van de kennis in het systeem.
- De toe te passen redeneerstrategieën.
- De organisatie van de kennis.
- De modulariteit van het systeem.
- Het modelleren van onzekerheid.

In het vervolg van dit artikel zal worden toegelicht hoe in de fase van probleemidentificatie en conceptualiseren deze keuzes meer expliciet gemaakt kunnen worden. Hierbij wordt gebruik gemaakt van het idee van generieke taken dat is ontwikkeld door Chandrasekaran (1988). Tenslotte zal de relevantie van deze fasen in de life-cycle van medische kennissystemen voor de acquisitie van medische kennis worden besproken.

Generieke taken

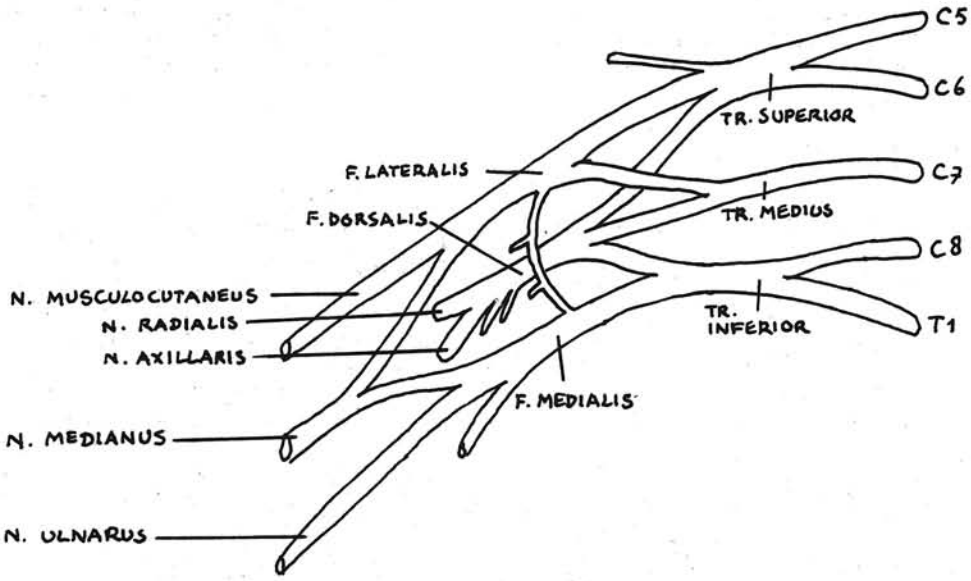
Chandrasekaran (1986, 1988) stelt dat het abstraktieniveau van de ondersteuning die wordt geboden bij het realiseren van kennissystemen te laag is. Deze ondersteuning bevindt zich op het niveau van de implementatie (het niveau van frames, produktieregels etc.), terwijl de problemen liggen op een hoger abstraktieniveau, dat van kennis en controle van het redeneerproces. Generieke taken bieden een hoger abstraktieniveau voor het oplossen van deze problemen. Zij kunnen worden opgevat als de bouwstenen waaruit complexe redeneerstrategieën kunnen worden opgebouwd. Iedere generieke taak wordt gekarakteriseerd door:
1. De informatie die nodig is als input voor de taak en de informatie die als gevolg van het uitvoeren van de taak wordt geproduceerd.
2. De wijze waarop kennis voor de taak dient te worden gerepresenteerd en georganiseerd.
3. Het proces van inferentie en controle dat de taak gebruikt.

Het realiseren van kennissystemen gebaseerd op generieke taken levert zodoende als vanzelfsprekend een oplossing voor een aantal van de genoemde problemen: te gebruiken redeneerstrategie, kennisorganisatie en modulariteit. Dit zal worden toegelicht met een voorbeeld van een neurologisch kennissysteem voor de diagnostiek van plexus brachialis letsels, PLEXUS (Jaspers, 1987).

De diagnostiek van plexus brachialis letsels

De plexus brachialis (Figuur 1) is een complexe zenuwstruktuur die zich bevindt in het overgangsgebied van de nek naar de bovenarm. Deze innerveert de zintuigen en spieren in de schoudergordel, de arm en de hand. Plexus brachialis letsels resulteren in geheel of gedeeltelijk funktieverlies van deze spieren en zintuigen. Vroegtijdige diagnostiek van plexus brachialis letsels is van groot belang voor de selektie van patiënten die in aanmerking komen voor neurochirurgische rekonstruktie. Tot 4 maanden na het trauma is neurochirurgie geïndiceerd, daarna wordt de prognose van deze behandeling veel slechter.

Retrospektief onderzoek onder 136 patiënten met plexus brachialis letsels toonde een behoefte aan beslissingsondersteuning bij diagnostiek en behandeling (Jaspers, 1986). In het vervolg zal de realisatie van een kennissysteem voor het lokaliseren van plexus brachialis letsels worden behandeld, uitgaande van de fasen in een kennissysteem life-cycle zoals die door Buchanan et al (1983) zijn geïntroduceerd.



Figuur 1: Schematische weergave van de plexus brachialis.

Probleemidentificatie en conceptualiseren

Traumatische plexus brachialis letsels bestaan doorgaans uit meervoudige letsels, waarbij de zenuwstruktuur in een groot gebied is beschadigd. Met name bij traktieletsels komen letsels op 2 of 3

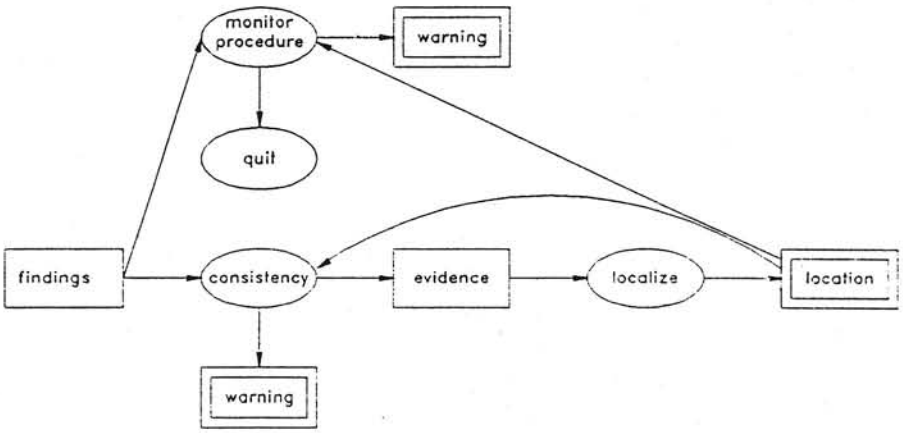
niveaus in de plexus brachialis voor. Door het grote aantal zenuwlokaties waaruit de plexus brachialis is opgebouwd en door het bestaan van meervoudige letsels is het aantal mogelijke kombinaties van letsels praktisch onbeperkt. Dit maakt het lokaliseren van plexus brachialis letsels bijzonder moeilijk. Daarnaast spelen met betrekking tot het diagnostiseren van deze letsels nog een aantal andere problemen een belangrijke rol:

1. Noodzakelijke diagnostische tests worden niet altijd uitgevoerd.
2. Sommige onderzoeken leveren slechts in specifieke gevallen geldige informatie op, bovendien worden onderzoeksresultaten dikwijls foutief geïnterpreteerd ten gevolge van anatomische variaties in de plexus brachialis.
3. Symptomen zijn vaak weinig specifiek, hetgeen het lokaliseren van plexus brachialis letsels verder bemoeilijkt.

Deze problemen suggereren dat een diagnostisch systeem voor het lokaliseren van plexus brachialis letsels de volgende taken dient uit te voeren (Figuur 2):

1. Controleer of de juiste diagnostische procedure wordt gevolgd ('monitor procedure').
2. Controleer de betrouwbaarheid en consistentie van onderzoeksresultaten ('consistency').
3. Lokaliseer het letsel ('localize').

In dit voorbeeld zal de taak 'localize' verder worden uitgewerkt.



Figuur 2: Taakstruktuur van een kennissysteem voor het lokaliseren van plexus brachialis letsels.

Niettegenstaande de genoemde problemen is de kennis betreffende het lokaliseren van plexus brachialis letsels vrij compleet. Een grote mate van onzekerheid wordt echter geïntroduceerd door de anatomische variaties en de niet-specificiteit van de symptomen, die het lokaliseren van deze letsels compliceren. Uitgaande van het gebruik van generieke taken in een kennissysteem voor het lokaliseren van plexus brachialis letsels, dient de taak 'localize' te worden opgesplitst in een aantal subtaken met gedefinieerde informatiein- en output, kennisorganisatie en inferentie en controle. Een eerste stap daartoe vormt het analyseren van het redeneerproces dat wordt gevolgd door menselijke experts bij het localiseren van plexus brachialis letsels. Dit proces is blootgelegd door het interviewen van deze experts.

Redeneerstrategie

De oplossingsruimte van het aantal plexus brachialis letsels bestaat uit ongeveer 2^{40} oplossingen. Uitputtend doorzoeken van deze ruimte voor het vinden van de oplossing is uiteraard uitgesloten. In (Jaspers, 1990) wordt aangetoond dat de redeneerstrategie die experts gebruiken voor het lokaliseren van plexus brachialis letsels achtereenvolgens bestaat uit een fase van data-abstraktie, een fase van inperken van de zoekruimte door middel van empirische associaties, een fase van verfijning van de gevonden oplossingen met behulp van meer diepe kennis van het proces en tenslotte een fase waarin uit de resterende hypothesen een definitieve oplossing (diagnose) wordt opgebouwd. Dit komt overeen met de strategie van 'heuristische klassifikatie' (Clancey, 1985), uitgebreid met een fase 'hypothese assemblage' die nodig is voor het construeren van een samengestelde oplossing bestaande uit meervoudige letsels, die het best de symptomen verklaart. Of in termen van de door Chandrasekaran (1988) gedefinieerde generieke taken bestaat de taak 'localize' uit:
- knowledge directed information passing.
- hypothesis matching.
- hierarchical classification.
- abductive hypothesis assembly.

Deze observatie legt het conceptuele model van het te realiseren kennissysteem vast. Bovendien is met het identificeren van de te implementeren redeneerstrategie ook een oplossing gevonden voor het organiseren van de kennis, daar elke generieke taak de wijze van organisatie voorschrijft. Ook het niveau van de te implementeren kennis (diep of oppervlakkig) is grotendeels gedefinieerd en de modulariteit van het kennissysteem is op twee niveaus gegarandeerd: Het kennissysteem valt uiteen in vier modules die ieder een generieke taak representeren. De interfaces tussen deze modules zijn gedefinieerd door de input en output van deze taken. Daarnaast biedt de taak 'hierachical classification' (Gomez en Chandrasekaran, 1981) nog een extra mogelijkheid voor het hierarchisch organiseren van de kennis.

Formaliseren en implementatie

Bij het formaliseren van het conceptuele model blijft een inzichtelijke organisatie van de oppervlakkige kennis in de module 'hypothesis match' en het modelleren van onzekerheid een probleem. In (Jaspers, 1987, 1990) wordt een methode gepresenteerd waarmee in het algemeen de onzekerheid in de 'evoking strength' van symptomen voor een hypothese expliciet en niet-numeriek kan worden gerepresenteerd. Deze zogenaamde 'classification of evidence' biedt bovendien een expliciete representatie voor de inferentiestruktuur van de 'hypothesis match' taak en een raamwerk voor de organisatie van de oppervlakkige kennis in deze taak. Met behulp van deze methode wordt de 'evoking strength' van geabstraheerde informatie voor iedere hypothese gerepresenteerd in categorieën. Voor het lokaliseren van plexus brachialis letsels bleken vijf van dergelijke categorieën noodzakelijk:

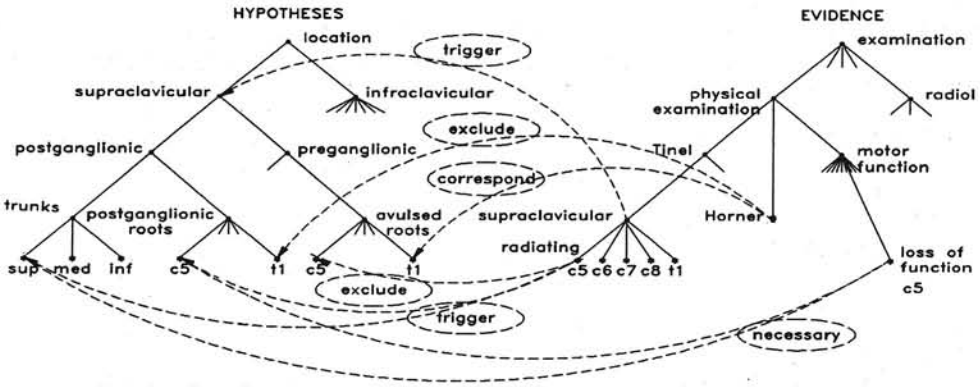
1. Triggering facts. Deze aktiveren en confirmeren een hypothese, ongeacht het bestaan van 'necessary' of 'exclusionary' feiten.
2. Necessary facts. Deze moeten voor een hypothese aanwezig zijn teneinde hem te kunnen postuleren.
3. Exclusionary facts. Het bestaan van een of meer van deze feiten verwerpt een hypothese.
4. Corresponding facts. Wanneer een specifieke hypothese is gepostuleerd worden deze feiten daardoor bevestigd.
5. Irrelevant facts. Voor een specifieke hypothese zijn sommige feiten irrelevant.

Met behulp van deze feiten wordt de redeneerstrategie van de 'hypothesis match' taak expliciet gedefinieerd. Deze strategie kan bijvoorbeeld worden gerepresenteerd in de vorm van vier soorten produktieregels (Jaspers, 1987):

1. Triggering rules. Deze aktiveren en confirmeren een hypothese met behulp van 'triggering facts'.
2. Pruning rules. Deze sluiten een hypothese uit met behulp van 'exclusionary facts'.
3. Evaluation rules. Deze evalueren een hypothese met behulp van 'necessary' en 'exclusionary facts'.
4. Confirmatory rules. Deze verklaren resterende symptomen op basis van de gepostuleerde hypothesen, met behulp van de 'corresponding facts'.

De 'classification of evidence' maakt het mogelijk oppervlakkige kennis op een inzichtelijke wijze te organiseren per hypothese. Daarnaast biedt het een expliciete modellering voor de onzekerheid in de 'evoking strength' van symptomen (Figuur 3). Uit de figuur blijkt dat een symptoom in staat kan zijn bepaalde hypothesen uit te sluiten, zonder dat het andere hypothesen confirmeert.

Met de opsplitsing van de 'localize' taak in vier generieke taken is een raamwerk gedefinieerd bestaande uit taken met bijbehorende



Figuur 3: 'Evoking strength' van een aantal feiten voor diverse hypothesen uit een klassifikatiehierarchie voor plexus brachialis letsels.

kennisorganisatie-strukturen en inferentiemechanismen. Dit raamwerk moet worden ingevuld met de specifieke kennis die voor ieder van de taken relevant is. Het totale diagnostische systeem van figuur 2 is gerealiseerd in 30 modules, geïmplementeerd in Delfi3 (Jonker, 1990). Het voert te ver deze hier te behandelen. Er wordt volstaan met een voorbeeld van de 'classification of evidence' uit de 'hypothesis match' taak (figuur 4).

Relevantie van het concept van 'generieke taken' voor de acquisitie van medische kennis.

Het voorgaande heeft aangetoond dat het identificeren van de te gebruiken redeneerstrategie voor een medische applicatie en het opsplitsen van deze strategie in generieke taken een handzaam raamwerk biedt voor het realiseren van medische kennissystemen. Met de benadering via generieke taken worden een aantal problemen op een hoog abstraktieniveau als vanzelfsprekend opgelost, door de wijze waarop generieke taken zijn gedefinieerd. Door het ontwikkelen van 'knowledge engineering toolboxes' bestaande uit generieke bouwstenen gebaseerd op generieke taken wordt het realiseren van medische kennissystemen aanzienlijk vereenvoudigd. Kennisorganisatie, inferentiemechanisme en input en output van deze taken zijn reeds gedefinieerd. Zoals getoond vergemakkelijken generieke taken de keuze voor het niveau van de te implementeren kennis, voor de organisatie van de kennis en voor het modulair opbouwen van medische kennissystemen. Daarnaast biedt deze aanpak en de 'classification of evidence' duidelijke ondersteuning bij de acquisitie van de benodigde kennis, omdat goed gedefinieerd is welk type kennis wordt gezocht. Dit maakt het mogelijk dit acquisitieproces veel gerichter te doen plaatsvinden.

```
DOBJ *Hypothesis
PRIVATE   triggers              : <Evidence> DEF No_evidence
        |  necessary            : <Evidence> DEF No_evidence
        |  exclusive            : <Evidence> DEF No_evidence
        |  corresponding        : <Evidence> DEF No_evidence
        |  asserted             : <BOOL> IFN [ deduce_hypothesis(triggers,
                                        necessary,exclusive,corresponding,
                                        'asserted)]
EOBJ

DOBJ *Evidence
PRIVATE present: <BOOL> DEF FALSE
EOBJ

IOBJ *No-evidence: Evidence(FALSE)
EOBJ

DREL deduce_hypothesis
DOMAIN    trig, nec, excl, corr    : <Evidence>
RANGE     asserted                 : <BOOL>
[[ [    trig.present
   OR   nec.present AND NOT excl.present
   ]    AND ! AND asserted = TRUE
OR asserted = FALSE
 ]
]
EREL
```

Figuur 4: Definitie van de 'classification of evidence' in Delfi3.

De resultaten van het gerealiseerde systeem PLEXUS tonen aan dat het goed mogelijk is op deze wijze medische kennissystemen te realiseren die aan de gestelde systeemeisen voldoen. PLEXUS is goed onderhoudbaar door zijn modulaire opzet, de kwaliteit van het advies is op het niveau van menselijke experts en het systeem is inzichtelijk door de wijze van kennisorganisatie en modelleren van onzekerheid (Jaspers, 1990).

Referenties

Blom J.A. (1975). Trend prediction and automated therapy in patient intensive care. In: Computers in cardiology, Rotterdam, pp. 213-214.

Buchanan B.G., Barstow D., Bechtel R. et al (1983). Constructing an expert system. In: Building expert systems, Hayes Roth F. et al eds., Reading, MA, Addison Wesley, pp. 127-167.

Chandrasekaran B. (1986). Generic tasks in knowledge-based reasoning: high-level building blocks for expert system design. IEEE Expert,

Fall 1986, pp. 23-30.

Chandrasekaran B. (1988). Generic tasks as building blocks for knowledge-based systems: the diagnosis and routine design examples. The Knowledge Engineering Review, vol. 3, pp. 183-210.

Clancey W.J. (1985). Heuristic classification. AI, vol. 27, pp. 289-350.

Daalen C. van (1988). Factors influencing medical expert system acceptance. Rapport WMR-N-284, Delft, TU-Delft, 47 p.

Fries J. (1972). Time-oriented patient records and a computer data bank. JAMA, vol. 222, pp. 1536-1542.

Gomez F., Chandrasekaran B. (1981). Knowledge organization and distribution for medical diagnosis. IEEE Trans. SMC, vol. SMC-11, pp. 34-42.

Gorry G.A., Barnett G.O. (1968). Experience with a model of sequential diagnosis. Comp.Biomed.Res., vol. 1, pp. 490-507.

Jaspers R.B.M. (1986). Diagnostiek van plexus brachialis letsels. Raport WMR-N-259, Delft, TU-Delft, 32 p.

Jaspers R.B.M., Helm F.C.T. van der (1987). Computer aided diagnosis and treatment of brachial plexus injuries. In: Lecture notes in medical informatics, vol. 33, Proc. AIME87, Fox J. et al eds., Berlijn, Springer Verlag, pp. 237-246.

Jaspers R.B.M. (1990). Medical decision support: an approach in the domain of brachial plexus injuries. Dissertatie, TU-Delft, 284 p.

Jonker W. (1990). The design and implementation of a knowledge representation and processing language. Dissertatie, RU-Utrecht, 289 p.

Neale I.M. (1988). First generation expert systems: a review of knowledge acquisition methodologies. The Knowledge Engineering Review, vol. 3, pp. 105-145.

Starmer C., Lee K., Harrell F., Rosati R. (1979). A database approach for stabilizing clinical decisions in the setting of chronic illness. Proc. Third SCAMC, IEEE, pp. 777-786.

Stassen H.G., Lunteren A. van, Hoogendoorn R. et al (1980). A computer model as an aid in the treatment of patients with injuries of the spinal cord. Proc. ICCS, Cambridge, MA, IEEE, pp. 385-390.

Warner H.R., Toronto A.F., Veasy L.G. (1964). Experience with Bayes' theorem for computer diagnosis of congenital heart disease. Ann.N.Y.Acad.Sc., vol. 115, no. 2.

Dit boek bevat de bijdragen van sprekers op de studiemidddag "Acquisitie van medische kennis ten behoeve van expertsystemen".
Aan de orde komen onder meer: de representatie van onzekerheid in kennis ten behoeve van medische expertsystemen, de invloed van biomedische en klinische kennis, theorie en praktijk op het gebied van de thallium-201 tomografie, de implementatie van anatomische kennis in een systeem voor de automatische labeling van bloedvaten in angiogrammen, de toepasbaarheid van technieken voor automatisch leren en de life-cycle van medische beslissingsondersteunende systemen.
De studiemiddag vond plaats op donderdag 3 mei 1990 aan de Technische Universiteit Delft en werd georganiseerd door de vakgroep Informatietheorie van de faculteit der Elektrotechniek, het Thoraxcentrum van de Erasmus Universiteit Rotterdam en de Stichting Centrum Medische Techniek.