Semi-Automatic Knowledge Acquisition of Fuzzy Symptom-Diagnosis Relationships with Hepatitides

Michael Schuerz[*], Klaus-Peter Adlassnig[*], Charles Lagor[†],
Barbara Schneider[‡], and Georg Grabner[§]

[*]Department of Medical Computer Sciences, Section on Medical Expert and Knowledge-Based Systems, University of Vienna Medical School, Spitalgasse 23, A-1090 Wien, Austria
Phone: +43-1-40400-6664, Fax: +43-1-40400-6667
email: {Michael.Schuerz | kpa}@akh-wien.ac.at
[†]Department of Medical Informatics, University of Utah, Salt Lake City, USA
email: Charles.Lagor@m.cc.utah.edu
[‡]Department of Medical Statistics and Documentation, University of Vienna Medical School, Schwarzspanierstrasse 17, A-1090 Wien, Austria
Phone: +43-1-4277-63207, Fax: +43-1-4277-9632
email: Barbara.Schneider@univie.ac.at
[§]Professor emeritus of the Second Department for Gastroenterology and Hepatology and of the Department of Medical Computer Sciences, University of Vienna Medical School, Austria
email: Georg.Grabner@teleweb.at

ABSTRACT: In this paper a method for forming fuzzy relationships between symptoms and diagnoses in the medical application area of hepatitides is being presented. These fuzzy relationships are defined for the frequency of occurrence of symptoms with diseases and the strength of confirmation of symptoms for diseases using relative sigma-counts. This process of knowledge acquisition is being carried out for constructing the knowledge base of the medical knowledge-based system CADIAG-II.

KEYWORDS: Knowledge Acquisition, Fuzzy Relations, Knowledge-Based Systems, Hepatitides, Medicine

INTRODUCTION

The medical expert system CADIAG-II infers diagnoses for a specific patient from symptoms using a knowledge base. These symptoms can be subjective complaints of the patient, findings of a clinical examination (= *signs*) or laboratory test results. Since medical data is immanently vague over wide ranges fuzzy sets are used to represent linguistic medical concepts such as, *reduced*, *normal*, *elevated*, and *highly elevated*. Such concepts are usually used by physicians to differentiate between different levels of strength concerning a specific symptom in the diagnostic and therapeutic process.

Although the frequency of occurrence of symptoms with diseases and the strength of confirmation of symptoms for diseases suggest an interpretation of relationships between symptoms and diagnoses as conditional probabilities either as frequency of occurrence or as strength of confirmation, the property of symptoms of having intermediate degrees of compatibility and the occurrence of a suspected diagnosis in a specific patient prevent a purely probabilistic calculation. (Adlassnig (1986a), p. 280)

Hence, for the relationships between symptoms and diagnoses fuzzy relations were formed as part of the knowledge base of CADIAG-II. These fuzzy relations are represented by proportions of cardinalities of fuzzy sets and are calculated using relative sigma-counts (Zadeh (1981), pp. 301–310) to obtain the frequency of occurrence of symptoms with diseases and the strength of confirmation of symptoms for diseases.

MATERIAL AND METHODS

For this retrospective study, 668 case records of adult hepatitis patients were selected from the online medical information system WAMIS (Grabner (1985), pp. 251–302). The patients' clinical diagnoses made are listed below; each clinical diagnosis was serologically verified and was thus considered to be a gold standard. The numbers in brackets are the appropriate ICD-9 codes:

- 36 patients with type A hepatitis (070.1),
- 114 patients with type B hepatitis (070.3),
- 22 patients with non A non B hepatitis (070.5),
- 269 patients with chronic hepatitis (571.4, 571.40, 571.41, 571.42, 571.48, 571.49),
- 27 patients with alcoholic hepatitis (571.1),
- 33 patients with hepatitis carriers (V02.6), and
- 167 patients with psychophysiologic disorders (306.9).

The patients in the last group did not present any liver diseases and hence they formed the *reference group* for this study.

The patients above received stationary treatment in the Vienna General Hospital (AKH-Wien) within the period from April, 1 1976 to March 31, 1986. Due to the given technical and organizational setting, more recent data could not be obtained. Consequently, it was not distinguished between hepatitis C, D, E, F, and G; the general diagnosis "hepatitis non A non B" was used instead. Based on the clinical picture and the prevalence of the different types of hepatitides in Vienna, most patients which had the diagnosis "hepatitis non A non B" would nowadays be regarded as hepatitis C cases.

The following parameters of each patient were investigated:

- alanine aminotransferase,
- alkaline phosphatase,
- aspartate aminotransferase,
- bilirubin,
- gamma-glutamyltranspeptidase,
- lactate dehydrogenase, and

the electrophoresis parameters

- albumin,
- alpha 1 globulin,
- alpha 2 globulin,
- beta globulin, and
- gamma globulin.

These parameters are important with patients presenting liver diseases. Based upon their patterns, they can give clinical evidence for the presence of a particular type of hepatitis.

From this data material *S*- and *p*-fuzzy membership functions were constructed semi-automatically for the linguistic, medical concepts—or linguistic variables—*reduced* ($m_{\downarrow}$), *normal* ($m_{\perp}$), *elevated* ($m_{\uparrow}$), and *highly elevated* ($m_{\uparrow\uparrow}$) for the *data-to-symbol conversion* of laboratory test results into symptoms. This was performed for all laboratory parameters except for the electrophoresis parameters albumin, alpha 1 globulin, and alpha 2 globulin (Schuerz (1998), pp. 23–30); for the latter three parameter values a calculation of a fuzzy membership function for *highly elevated* need not to be considered since such a fine distinction is usually not used by physicians.

Then fuzzy relations between symptoms represented by the above-mentioned fuzzy sets and diagnoses were calculated for the frequency of occurrence of symptoms with diseases (m_O) and the strength of confirmation of symptoms for diseases (m_C).

More formally, a fuzzy relation $\tilde{R}$ from a set of fuzzy symptoms $S = \{\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_l\}$ into a set of diagnoses $D = \{\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_m\}$ is defined as the mapping (function) of the cartesian product $S \times D$ into the closed real interval $[0,1]$. Then the fuzzy relation $\tilde{R}$ denotes the fuzzy degree of relationship between $\tilde{s}_i \in S, 1 \le i \le n,$ and $\tilde{d}_j \in D, 1 \le j \le m,$ for every ordered pair $\langle \tilde{d}_i, \tilde{s}_j \rangle \in \tilde{S} \times D$. Clearly, $\tilde{R}$ is a fuzzy set $\tilde{S} \times D$ into $[0,1]$ (cf. Kerre (1991), pp. 86–88)

To obtain relationships between symptoms and diagnoses relative sigma-counts were used to represent these fuzzy relations at this medical area of application, i. e.,

$$m_O\left(\tilde{s}_i,\tilde{d}_j\right) := \sum Count\left(\tilde{s}_i/\tilde{d}_j\right) = \frac{\sum Count\left(\tilde{s}_i \cap \tilde{d}_j\right)}{\sum Count\left(\tilde{d}_j\right)} = \frac{\sum_{k=1}^{n} \min\left\{m_{PS}\left(p_k,\tilde{s}_i\right), m_{PD}\left(p_k,\tilde{d}_j\right)\right\}}{\sum_{k=1}^{n} m_{PD}\left(p_k,\tilde{d}_j\right)}$$

and

$$m_C\left(\tilde{s}_i,\tilde{d}_j\right) := \sum Count\left(\tilde{d}_j/\tilde{s}_i\right) = \frac{\sum Count\left(\tilde{s}_i \cap \tilde{d}_j\right)}{\sum Count\left(\tilde{s}_i\right)} = \frac{\sum_{k=1}^{n} \min\left\{m_{PS}\left(p_k,\tilde{s}_i\right), m_{PD}\left(p_k,\tilde{d}_j\right)\right\}}{\sum_{k=1}^{n} m_{PD}\left(p_k,\tilde{s}_i\right)}$$

where m_{PS} denotes the fuzzy relation from a set of patients $P = \{p_1, p_2, \ldots, p_n\}$ into a fuzzy set of symptoms $S = \{\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_l\}$ and m_{PD} is the fuzzy relation between the set of patients *P* and the set of diagnoses *D*.

Since the occurrence of a disease in a collective of patients, the so-called *prevalence*, $P\left(\tilde{d}_j\right)$ is—in general—different in different collectives of patients the strength of confirmation of a symptom for a disease can also be taken into account by prevalences which are assumed to be equal (m_C^*). This means that during the consultation process with a medical expert system using this mechanism diagnostic proposals for frequent as well as for rare diseases would be inferred with equal possibility. In detail

$$\sum Count\left(\tilde{d}_j\right) = \sum Count\left(\bar{\tilde{d}}_j\right) := 0.5$$

covers this approach for the sigma-count extension of conditional probabilities. Hence,

$$m_C^*\left(\tilde{s}_i,\tilde{d}_j\right) := \sum Count\left(\tilde{d}_j/\tilde{s}_i\right) = \frac{\sum Count\left(\tilde{d}_j \cap \tilde{s}_i\right)}{\sum Count\left(\tilde{s}_i\right)} = \frac{\sum Count\left(\tilde{d}_j\right) \cdot \sum Count\left(\tilde{s}_i/\tilde{d}_j\right)}{\sum Count\left(\tilde{s}_i\right)} =$$

$$= \frac{\sum Count\left(\tilde{d}_j\right) \cdot \sum Count\left(\tilde{s}_i/\tilde{d}_j\right)}{\sum Count\left(\tilde{d}_j\right) \cdot \sum Count\left(\tilde{s}_i/d_j\right) + \sum Count\left(\bar{\tilde{d}}_j\right) \cdot \sum Count\left(\tilde{s}_i/\bar{\tilde{d}}_j\right)} = \frac{\sum Count\left(\tilde{s}_i/\tilde{d}_j\right)}{\sum Count\left(\tilde{s}_i/\tilde{d}_j\right) + \sum Count\left(\tilde{s}_i/\bar{\tilde{d}}_j\right)}$$

with

$$\sum Count\left(\tilde{s}_i/\bar{\tilde{d}}_j\right) = \frac{\sum Count\left(\tilde{s}_i \cap \bar{\tilde{d}}_j\right)}{\sum Count\left(\bar{\tilde{d}}_j\right)} = \frac{\sum_{k=1}^{n} \min\left\{m_{PS}\left(p_k,\tilde{s}_i\right), m_{P\bar{D}}\left(p_k,\bar{\tilde{d}}_j\right)\right\}}{\sum_{k=1}^{n} m_{P\bar{D}}\left(p_k,\bar{\tilde{d}}_j\right)}.$$

RESULTS

Fuzzy relations were calculated for the frequency of occurrence of symptoms with diseases (m_O), the strength of confirmation of symptoms for diseases including the prevalences of diseases in the used data material (m_C), and the strength of confirmation of symptoms for diseases by prevalences which are assumed to be equal (m_C^*) for all relationships between the corresponding symptoms and diagnoses. Table I shows these fuzzy relations on the example of type A hepatitis, type B hepatitis, and the reference group.

	type A hepatitis			type B hepatitis			reference group		
	μ_O	μ_C	μ_C^*	μ_O	μ_C	μ_C^*	μ_O	μ_C	μ_C^*
alanine aminotransferase									
reduced	0.00	0.00	0.00	0.00	0.00	0.00	0.02	1.00	1.00
normal	0.03	0.00	0.07	0.03	0.02	0.07	0.97	0.68	0.87
elevated	0.97	0.08	0.61	0.97	0.25	0.63	0.02	0.01	0.02
highly elevated	0.56	0.28	0.87	0.44	0.69	0.92	0.00	0.00	0.00
alkaline phosphatase									
reduced	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.70	0.87
normal	0.31	0.02	0.27	0.55	0.11	0.40	0.97	0.31	0.57
elevated	0.69	0.16	0.79	0.45	0.35	0.73	0.02	0.02	0.06
highly elevated	0.03	0.05	0.51	0.05	0.27	0.65	0.00	0.00	0.00
aspartate aminotransferase									
reduced	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.91	0.97
normal	0.09	0.01	0.15	0.09	0.03	0.15	0.98	0.54	0.78
elevated	0.91	0.09	0.64	0.91	0.29	0.66	0.01	0.00	0.01
highly elevated	0.29	0.17	0.78	0.40	0.77	0.94	0.00	0.00	0.00
bilirubin									
reduced	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.91	0.97
normal	0.09	0.01	0.11	0.24	0.06	0.24	0.97	0.36	0.62
elevated	0.91	0.15	0.76	0.76	0.40	0.77	0.02	0.02	0.05
highly elevated	0.52	0.24	0.85	0.45	0.66	0.91	0.00	0.00	0.00
gamma-glutamyltranspeptidase									
reduced	0.00	0.00	0.00	0.01	1.00	1.00	0.00	0.00	0.00
normal	0.21	0.02	0.24	0.41	0.12	0.37	0.99	0.31	0.64
elevated	0.79	0.17	0.71	0.58	0.30	0.66	0.01	0.01	0.03
highly elevated	0.00	0.00	0.00	0.02	0.22	0.56	0.00	0.00	0.00
lactate dehydrogenase									
reduced	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.41	0.67
normal	0.57	0.03	0.41	0.40	0.08	0.31	0.96	0.31	0.56
elevated	0.43	0.11	0.71	0.60	0.51	0.84	0.02	0.03	0.07
highly elevated	0.16	0.27	0.88	0.11	0.62	0.89	0.00	0.00	0.00
albumin									
reduced	0.16	0.08	0.60	0.12	0.20	0.53	0.01	0.02	0.05
normal	0.84	0.05	0.49	0.88	0.19	0.50	0.98	0.28	0.54
elevated	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.69	0.87
alpha 1 globulin									
reduced	0.02	0.13	0.72	0.00	0.03	0.12	0.01	0.31	0.57
normal	0.97	0.05	0.50	1.00	0.19	0.50	0.98	0.25	0.50
elevated	0.01	0.08	0.60	0.00	0.00	0.00	0.01	0.43	0.69
alpha 2 globulin									
reduced	0.06	0.07	0.57	0.11	0.00	0.75	0.02	0.08	0.21
normal	0.94	0.06	0.50	0.88	0.00	0.48	0.97	0.26	0.51
elevated	0.00	0.00	0.02	0.01	0.00	0.41	0.02	0.33	0.59
beta globulin									
reduced	0.21	0.17	0.78	0.04	0.00	0.33	0.01	0.05	0.13
normal	0.73	0.05	0.45	0.90	0.00	0.51	0.96	0.28	0.53
elevated	0.06	0.07	0.56	0.06	0.00	0.59	0.02	0.12	0.28
highly elevated	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
gamma globulin									
reduced	0.00	0.00	0.00	0.01	0.00	0.59	0.02	0.48	0.73
normal	0.40	0.03	0.35	0.64	0.00	0.47	0.96	0.34	0.61
elevated	0.60	0.12	0.70	0.35	0.00	0.57	0.02	0.02	0.06
highly elevated	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table I:. Fuzzy relations between eleven selected liver laboratory parameter values with type A hepatitis, type B hepatitis, and the reference group.

DISCUSSION

The aim of this study was to develop a feasible method for semi-automatic knowledge acquisition of fuzzy symptom-diagnose relationships. This could be shown with the proposed method on the applied material. Since it cannot be excluded that a patient database contains only symptoms definitely present or absent, indicated by degrees of compatibility $m_{PS}(p_k, \tilde{s}_i) \in \{0,1\}$, and diagnoses definitely confirmed or excluded, expressed through $m_{PD}(p_k, \tilde{d}_j) \in \{0,1\}$, that is, if there is no uncertainty about the assignment of measured laboratory test results into linguistic medical concepts such as "bilirubin in serum *highly elevated*" and if there is no uncertainty about the clinical diagnoses of the considered patients, then the above-mentioned calculations coincide with the calculations for $P(\tilde{s}_i / \tilde{d}_j)$ and $P(\tilde{d}_j / \tilde{s}_i)$ (Adlassnig (1986b), pp. 213–214), analogous $m_{P\overline{D}}(p_k, \tilde{\tilde{d}}_j) \in \{0,1\}$. Since the fuzzy relations are calculated from sample data and hence may contain biases those fuzzy relations equal zero as well as those equal one need to be checked by a physician according to their frequency of occurrence respectively strength of confirmation in the parent population. The obtained fuzzy relations which are part of the knowledge base of CADIAG-II are applied during the inference process to derive diagnoses from symptoms concerning a specific patient. Because this resulting diagnostic proposals are made in the context of a specific patient and not concerning a collective of patients at the consultation process the strength of confirmation of symptoms for diseases by prevalences which are assumed to be equal yields a better approach to point the applying physician also to seldom diseases with usually less prevalences in the local geographical area of application. Another crucial aspect is the definition of the fuzzy sets for *elevated* and *highly elevated*. The proposed method defines *highly elevated* to be a proper fuzzy subset of *elevated*, a so-called *non-complementary fuzzy set*. Such a relation between fuzzy sets is used if the considered fuzzy sets represent gradual differentiations within the same diagnosis. The symptoms may have a different strength of confirmation for such a disease. On the other hand definitions of fuzzy sets for *elevated* and *highly elevated* would also be possible in a way that the function value of $m_{\downarrow}$ decreases the same amount the function value of $m_{\uparrow\uparrow}$ increases on its left-hand edge (= *complementary fuzzy set*). In this case the corresponding fuzzy sets would not represent gradual differentiations within the same diagnosis but allow to differentiate among different diagnoses (Leitich (1995) p. 22, Bögl (1997), p. 85–86). This method will remain semi-automatic because the results have to be checked for plausibility and—should the occasion arise—adapted by a physician for the intended application. Future work will cover not only an evaluation of different approaches concerning the fuzzy sets for forming fuzzy relations but also an application of this method to other medical areas.

REFERENCES

ADLASSNIG, KLAUS-PETER; SCHEITHAUER, WERNER; KOLARZ, GERNOT, 1986a, "Fuzzy Medical Diagnosis in a Hospital", in: Negoita, Constantin V.; Prade, Henri, "Fuzzy logic in knowledge engineering", Köln: Verlag TÜV Rheinland, pp. 275–294.

ADLASSNIG, KLAUS-PETER; KOLARZ, GERNOT; SCHEITHAUER, WERNER; GRABNER, HELMUT, 1986b, "Apporach to a hospital-based application of a medical expert system", Med. Inform., vol. 11, no. 3, pp. 205–223.

BÖGL, KARL, 1997, "Design and Implementation of a Web-Based Knowledge Acquisition Toolkit for Medical Expert Consultation Systems", Ph. D. Thesis, Vienna: Vienna University of Technology.

GRABNER, GEORG, ed., 1985, "WAMIS – Wiener Allgemeines Medizinisches Informations-System", Heidelberg–Berlin–New York: Springer-Verlag.

KERRE, ETIENNE, E., 1991, "Introduction to the Basic Principles of Fuzzy Set Theory and some of its Applications", Gent: Communication & Cognition.

LEITICH, HARALD, 1995, "Anforderungen an ein Wissenserwerbssystem für das medizinische Expertensystem CADIAG-4", Master's Thesis, Vienna: Vienna University of Technoloy.

SCHUERZ, MICHAEL; ADLASSNIG, KLAUS-PETER; LAGOR, CHARLES; SCHNEIDER, BARBARA; GRABNER, GEORG, 1998, "Supervised Construction of Fuzzy Sets from Sample Data: Laboratory Test Results Describing Different Forms of Hepatitis", Workshop on Applications of Fuzzy Logic, Vienna, Austria, pp. 23–30.

ZADEH, LOTFI A., 1981, "Test-score semantics for natural languages and meaning representation via PRUF", Technical Note 247, Menlo Park/CA: AI Center, SRI International (also in: Rieger, Burghard, B., ed., "Empirical Semantics", Bochum: Brockmayer, pp. 281–349)