ELSEVIER Artificial Intelligence in Medicine 14 (1998) 101–117

Artificial Intelligence in Medicine

Acquiring background knowledge for machine learning using function decomposition: a case study in rheumatology

Blaž Zupan *, Sašo Džeroski

Department of Intelligent Systems, Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia

Abstract

Domain or background knowledge is often needed in order to solve difficult problems of learning medical diagnostic rules. Earlier experiments have demonstrated the utility of background knowledge when learning rules for early diagnosis of rheumatic diseases. A particular form of background knowledge comprising typical co-occurrences of several groups of attributes was provided by a medical expert. This paper explores the possibility of automating the process of acquiring background knowledge of this kind and studies the utility of such methods in the problem domain of rheumatic diseases. A method based on function decomposition is proposed that identifies typical co-occurrences for a given set of attributes. The method is evaluated by comparing the typical co-occurrences it identifies as well as their contribution to the performance of machine learning algorithms, to the ones provided by a medical expert.

Keywords: Background knowledge; Knowledge acquisition and validation; Inductive learning; Typical co-occurrences; Function decomposition; Diagnosis of rheumatic diseases

* Corresponding author. Tel.: +386 61 1773380; fax: +386 61 1251038; e-mail: blaz.zupan@ijs.si

0933-3657/98/$19.00
PII S0933-3657(98)00018-9

1. Introduction

When using machine learning to learn medical diagnostic rules from patient records, it may be desirable to augment the latter with additional diagnostic knowledge about the particular domain, especially for difficult diagnostic problems. In machine learning terminology, additional expert knowledge is usually referred to as *background knowledge*.

A particular type of expert knowledge specifies which combinations of values *(co-occurrences)* of a set *(grouping)* of attributes have high importance for the classification problem at hand. These combinations of values are called *typical co-occurrences*. An expert would specify the groupings as well as the typical co-occurrences associated with them.

This paper proposes a technique for computer supported acquisition of typical co-occurrences and studies its utility in the difficult problem of early diagnosis of rheumatic diseases [9,14,15]. In this domain, the task is to diagnose patients into one of eight diagnostic classes, given 16 anamnestic attributes. The difficulty of the diagnostic problem itself and noise in the data make this a very difficult problem for machine learning approaches. A more detailed description of the domain can be found in Section 3. When asked for some additional knowledge about this domain, a medical expert proposed six groupings (pairs or triples of attributes) and their typical co-occurrences (characteristic combinations of values). These are given in Table 4 in Section 3. For each grouping, a new attribute is introduced and added to the data set, thus resulting in an extended data set to be considered in the learning process. For a particular patient record (example), this attribute has, as a value, the typical co-occurrence observed for the patient, if one was indeed observed, or has the value 'irrelevant' otherwise. A rule induction system, such as CN2 [5], or any attribute-value learning system can subsequently be applied to the extended learning problem.

To illustrate the concept, let us consider Grouping 2. It relates the attributes 'spinal pain' and 'duration of morning stiffness' where the typical co-occurrences are: no spinal pain and morning stiffness up to 1 h, spondylotic pain and morning stiffness up to 1 h, spondylitic pain and morning stiffness longer than 1 h. An example rule that uses this grouping and the second co-occurrence as induced from data by CN2 is given in Table 1.

The background knowledge in the form of typical co-occurrences was shown to have a positive effect on rule induction for early diagnosis of rheumatic diseases in

Table 1
A rule that makes use of a typical co-occurrence in the domain of diagnosis of rheumatic diseases

IF	Duration_of_present_symptoms >6.5 months
AND	Duration_of_rheumatic_diseases <5.5 years
AND	Number_of_painful_joints >16
AND	grouping2(Spinal_pain,Duration_of_morning_stiffness) = 'spondylotic & up to 1 h'
THEN	Diagnosis = Degenerative_spine_diseases

several respects. First, rules induced in the presence of background knowledge perform better in terms of classification accuracy and information content [15]. Second, it substantially improves the quality of induced rules from a medical point of view as assessed by a medical expert [15]. Finally, it reduces the effects of noisy data on the process of rule induction and nearest neighbor classification [9].

The described improvement in performance motivates the use of typical co-occurrences when learning medical diagnostic rules. Although the case described above is, to the best of our knowledge, the only domain for which the use of typical co-occurrences was documented, attribute groupings are frequently used within the well-established area of hierarchical decision support systems [20]. The motivation shared by the two approaches is that of 'divide-and-conquer': groupings usually include only a few attributes and can be regarded as a subproblem which is sufficiently simple for the expert to express the underlying relationships. In the multi-attribute decision support system DEX [3], such relationships are expressed as a tabulated function that generates the value of a new (intermediate) attribute given the values of the attributes in the grouping. Such a function is tabulated by the domain expert, who usually defines the value of the new attribute only for the most representative (typical) combination of attributes' values. DEX has been successfully applied in more than 50 realistic decision making domains. A similar technique called structured induction was proposed by Shapiro and Niblett [23]. Michie [16] reports on many successful industrial applications of structured induction. These two techniques differ from typical co-occurrences in that each co-occurrence maps to a distinct value of the new attribute, while different combinations of attributes' values may share the corresponding value of the new attribute in DEX and structured induction.

Before proceeding further, let us briefly mention other related work. The domain of early diagnosis of rheumatic diseases has been first treated with a machine learning approach by Pirnat et al. [19]. Decision tree based approaches have been further applied to this domain by Karalič and Pirnat [12]. The use of typical co-occurrences in this domain has been investigated by Lavrač et al. in combination with a decision tree approach [14] and in combination with a rule induction approach [15] and by Džeroski and Lavrač [9] in combination with nearest neighbor classification.

The drawback of the approach by Lavrač et al. [14] is that it relies on typical co-occurrences that have to be solicited from a medical expert. It is well-known that direct knowledge acquisition from experts is an arduous and error-prone process [11]. To provide computer support for this process, the paper proposes a method for automated acquisition of background knowledge in the form of typical co-occurrences. The expert need only specify the groupings, while the associated co-occurrences are determined automatically. A method that can identify most relevant groupings is also described.

The typical co-occurrence acquisition method proposed in this paper uses several fundamental algorithms from function decomposition. The pioneers of this field are Ashenhurst [1] and Curtis [7]. They have used function decomposition for the discovery of Boolean functions. Its potential use within machine learning was first

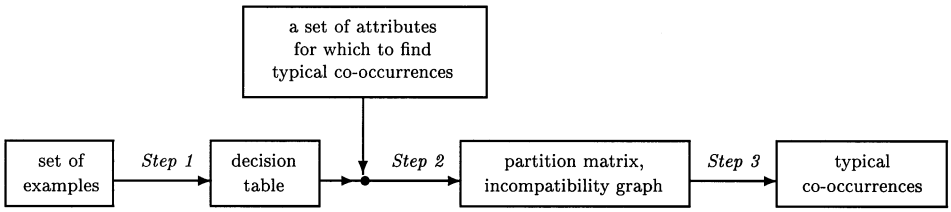

Fig. 1. The entities used and derived by the typical co-occurrence derivation method.

observed by Samuel [21] and Biermann [2]. A recent report of Perkowski et al. [18] provides a comprehensive survey of the literature on function decomposition. In this paper, we refer to the decomposition algorithms which use decision tables with multi-valued attributes and classes which were developed by Zupan et al. [25] and Bohanec et al. [4].

The remainder of the paper is organized as follows. Section 2 describes the method for acquisition of typical co-occurrences. Section 3 describes the domain of early diagnosis of rheumatic diseases, and the background knowledge provided by the expert. Taking the groupings provided by the expert, we apply the proposed method to determine the typical co-occurrences. The results of these experiments are also discussed in Section 3. Section 4 proposes a method for assisting the domain expert in selecting attribute groupings. Section 5 concludes and outlines some directions for further work.

2. The method

This section introduces, both formally and through an example, the method that derives typical co-occurrences for a given set of attributes from a given set of examples represented as attribute-value vectors with assigned classes. The overall data-flow of the method is shown in Fig. 1. The method first converts the set of examples to a decision table (Step 1). Next, decision table decomposition methods are used to derive a so-called partition matrix (Step 2). Finally, the typical co-occurrences for a given set of attributes are derived (Step 3), using an approach based on coloring the incompatibility graph of the partition matrix.

We first give an example of decision table decomposition and introduce the required decomposition methodology. The description of the method to acquire a set of typical co-occurrences is given next. For machine learning in medical domains, the data is usually represented as a set of examples, and we propose a technique to convert this representation to a decision table, a representation required by the proposed method. The section concludes with a brief note about the implementation.

2.1. Decision table decomposition: an example

Suppose we are given a *decision table* $y = F(X) = F(x_1,x_2,x_3)$ (Table 2) with three

Table 2
An example decision table

x_1	x_2	x_3	y
lo	lo	lo	lo
lo	med	hi	med
lo	hi	lo	lo
lo	hi	hi	hi
med	med	lo	med
med	hi	lo	med
med	hi	hi	hi
hi	lo	lo	hi
hi	hi	lo	hi

attributes x_1, x_2, and x_3, and class y, and we want to decompose it to decision tables G and H, such that $y = G(x_1,c)$ and $c = H(x_2, x_3)$. For this decomposition, an initial set of attributes X is partitioned to a *bound set* $\{x_2, x_3\}$ used with H and a *free set* $\{x_1\}$ used with G. Decomposition requires the introduction of a new attribute c which depends only on the variables in the bound set.

To derive G and H from F, we first need to represent a decision table with a *partition matrix* (Table 3). A partition matrix uses all possible combinations of attribute values from the bound set as column labels and those from the free set as row labels. Each column in a partition matrix specifies a behavior of the function F when the attributes in the bound set are constant. Two elements of a partition matrix are compatible if they are the same or at least one of them is unknown (denoted by '–'). Two columns are compatible if all of their elements are pairwise compatible: these columns are considered to represent the same behavior as function F.

The problem is now to assign labels to the columns of the partition matrix so that only groups of mutually compatible columns have the same label. Columns with the same label exhibit the same behavior in respect to F and can use a single value of the new concept c. Label assignment involves the construction of a *column*

Table 3
Partition matrix for the decision table from Table 2 with free set $\{x_1\}$ and bound set $\{x_2, x_3\}$

	x_2	lo	lo	med	med	hi	hi
x_1	x_3	lo	hi	lo	hi	lo	hi
lo		lo	—	—	med	lo	hi
med		—	—	med	—	med	hi
hi		hi	—	—	—	hi	—
color		3	3	3	2	3	1

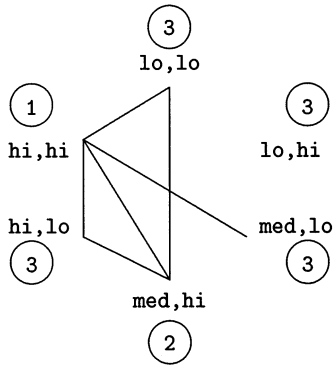

Fig. 2. Incompatibility graph for the partition matrix in Table 3.

incompatibility graph, where columns of the partition matrix are nodes and two nodes are connected if they are incompatible. Column labels are then assigned by coloring the incompatibility graph. For our example, the incompatibility graph with one of the possible optimal colorings is given in Fig. 2.

For better comprehensibility, we interpret the column labels (colors) as follows: '1' as `hi`, '2' as `med`, and '3' as `lo`. These labels and the partition matrix straightforwardly determine the function $c = H(x_2, x_3)$. To determine the function $G(x_1, c)$, we look up the annotated partition matrix for all the possible combinations of x_1 and c. The final result of the decomposition is represented as a hierarchy of two decision tables in Fig. 3. If we further examine the discovered functions G and H, we can see that $G \subset \text{MAX}$ and $H \subset \text{MIN}$.

2.2. Acquiring typical co-occurrences from a decision table

In the above example, different colors can be assigned to the same column of a partition matrix while retaining the minimal number of colors. For example, the column (`med, lo`) could be assigned either color 2 or 3, and the column (`lo, hi`) could be assigned any of the three colors used. On the other hand, the column (`lo, lo`) could only be assigned a single color due to the incompatibilities with (`med, hi`) and (`hi, hi`) which are assigned different colors. While only one distinct behavior exists for (`lo, lo`) with respect to F, there exist several for (`med, lo`) and (`lo, hi`). The combination (`lo, lo`) of attributes x_2 and x_3 thus tells us more about the behavior of function F and is therefore more typical. Moreover, the columns that can be assigned only one color form a foundation for such color assignments and will be called *typical columns* of the partition matrix *(typical nodes* of the incompatibility graph) and will further indicate for *typical co-occurrences* of the attributes in the bound set.

Therefore, for a given set of attributes for which we want to derive the typical co-occurrences (bound set) and for a given decision table, we have to first derive a corresponding partition matrix and its incompatibility graph. The algorithms for the construction of the partition matrix and incompatibility graph are described in

detail in [25]. The typical co-occurrences derivation method subsequently uses the incompatibility graph and discovers the typical co-occurrences through coloring. Since graph coloring is an NP-hard problem, the computation time of an exhaustive search algorithm is prohibitive even for small graphs with, say, 15 nodes. Instead, we use the Color Influence Method of polynomial complexity [18]. The Color Influence Method sorts the nodes to color by decreasing connectivity and then assigns to each node a color that is different from the colors of its neighbors so that a minimal number of colors is used. In this way, the coloring can have a single or several candidate colors for each node. The number of these candidate colors is used to determine the typicality of the node. We use the following definition:

Definition (Typical node n **of incompatibility graph** IG) A node n∈IG is typical if and only if, in the process of coloring using the Color Influence Method, it has only one candidate color to be assigned to.

The above definition is used to extend the Color Influence Method to both color the incompatibility graph and discover typical co-occurrences (Algorithm 1) at the same time.

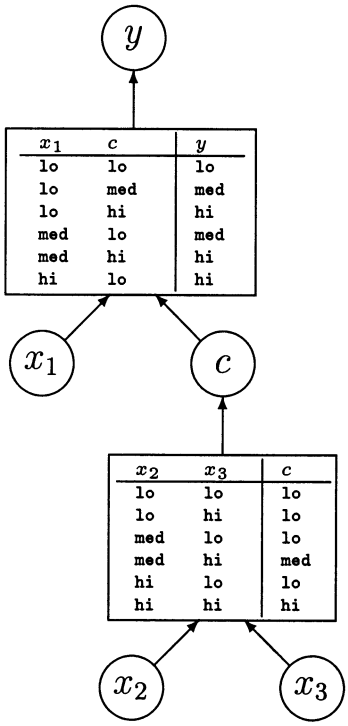

Fig. 3. The result of decomposing the decision table from Table 2.

Algorithm 1: Coloring of an incompatibility graph and selection of typical nodes

Input: incompatibility matrix IG
Output: typical co-occurrences for attributes in bound set

while there are no uncolored nodes in IG **do**
  select the uncolored node $n \in$ IG with highest connectivity
  if there are no colored non-adjacent nodes
    or all colored non-adjacent nodes have the same color
  then *n is* typical **else** *n is* not typical **endif**
  color *n* with the first free color different from the colors of adjacent nodes
endwhile

Let us illustrate the use of Algorithm 1 on the incompatibility graph from Fig. 2. The nodes sorted by decreasing connectivity are

(`hi, hi`), (`med, hi`), (`lo, lo`), (`hi, lo`), (`med, lo`), (`lo, hi`)

First, node (`hi, hi`) is selected, determined to be typical (no other nodes have been colored yet), and assigned color 1. Next, node (`med, hi`) is considered. There are no colored nodes non-adjacent to it and as a result, this node is typical. Since the adjacent node (`hi, hi`) has color 1, color 2 is assigned to (`med, hi`). Similarly, (`lo, lo`) is also typical and colored with color 3 because colors 1 and 2 have already been used for adjacent nodes (`hi, hi`) and (`med, hi`). Next, node (`hi, lo`) has a single colored non-adjacent node (`lo, lo`) and is thus typical and colored with the same color 3. The first nontypical node is (`med, lo`): it has three nodes (`med, hi`), (`lo, lo`), and (`hi, lo`) that are non-adjacent to it and use different colors 2 and 3. Among these, color 3 is then arbitrarily chosen for (`med, lo`). Similarly, node (`lo, hi`) is found not to be typical and among three candidate colors, color 3 is arbitrarily assigned to it. Therefore, among six possible combinations of attribute values, the algorithm found four typical co-occurrences: (`hi, hi`), (`med, hi`), (`lo, lo`), and (`hi, lo`).

The described method finds a possible set of typical nodes, but it does not guarantee that this is the only set of its kind. An alternative method that would search more exhaustively and possibly evaluate all different colorings of the incompatibility graph may be more complete and propose a different set of typical co-occurrences. However, its (possibly exponential) complexity would limit its applicability.

2.3. Derivation of a decision table from a set of examples

The typical co-occurrence derivation method requires domain data in the form of a decision table. Decision tables require nominal attributes, and for a specific combination of attribute values, define at most one class. However, data sets from medical domains often include continuous attributes and may contain several

examples with the same attribute values but different classes. Therefore, we need a method that, given a set of domain examples, would derive a corresponding decision table. For all continuous attributes, we assume that a discretization is given or can be derived from the examples.

The method is given in Algorithm 2. It searches through the set of examples E whose attribute values are the same if nominal or discretize to the same value if continuous. For such sets of examples E', a majority class value is found and a corresponding entry is added to the decision table. The examples from E' are then removed from E and the process is repeated until there are no more examples in E.

Algorithm 2: Derivation of a decision table from a set of examples

---

Input: Set of examples $E = \{e_i\}$, Discretization for continuous attributes
Output: Decision table DT

while $E \neq \varnothing$
  select $e_j \in E$
  find $E' = \{e_k;\ e_k \in E\}$ such that
    (1) for all discrete attributes, e_k has the same value as e_j
    (2) for all continuous attributes, e_k's discretized value is the same as e_j's
  $E' \leftarrow E' \cup \{e_j\}$
  $c \leftarrow$ a majority class value of the examples in E'
  add e_j with discretized continuous values and with class c to DT
  $E \leftarrow E \backslash E'$
endwhile

2.4. Implementation

The typical co-occurrences extraction method was implemented as $\mathsf{HINT}_{\mathsf{TCO}}$, an extension of the Hierarchy Induction Tool HINT [25] for learning concept hierarchies from examples by decision table decomposition. Both HINT and $\mathsf{HINT}_{\mathsf{TCO}}$ run on a variety of UNIX platforms, including HP/UX, SunOS and IRIX.

3. Identifying typical co-occurrences in the early diagnosis of rheumatic diseases

3.1. The domain

The data on early diagnosis of rheumatic diseases used in our experiments originate from the University Medical Center in Ljubljana [19] and comprise records on 462 patients. The multitude of over 200 different diagnoses have been grouped into three, six, eight or 12 diagnostic classes. Our study uses eight diagnostic classes: degenerative spine diseases, degenerative joint diseases, inflam-

matory spine diseases, other inflammatory diseases, extraarticular rheumatism, crystal-induced synovitis, non-specific rheumatic manifestations, and non-rheumatic diseases.

For each patient, 16 anamnestic attributes are recorded: sex, age, family anamnesis, duration of present symptoms (in weeks), duration of rheumatic diseases (in weeks), joint pain (arthrotic, arthritic), number of painful joints, number of swollen joints, spinal pain (spondylotic, spondylitic), other pain (headache, pain in muscles, thorax, abdomen, heels), duration of morning stiffness (in hours), skin manifestations, mucosal manifestations, eye manifestations, other manifestations, and therapy. The continuous attributes (age, durations and numbers of joints) have been discretized according to expert suggestions. For the continuous attributes that appear in groupings, the discretizations can be read out from Table 4. For example, from Table 4 we can see that the attribute 'duration of morning stiffness' has been discretized into two intervals: up to 1 h and longer than 1 h.

3.2. The background knowledge

In an earlier study [14], a specialist for rheumatic diseases provided his knowledge about typical co-occurrences of six groupings of attributes. The groupings and the co-occurrences are given in Table 4, where a bullet in the column marked 'specialist' and the row marked X means that tuple X is a typical co-occurrence for the corresponding grouping. For example, grouping 1 relates the attributes 'joint pain' and 'duration of morning stiffness', with typical co-occurrences defined by the expert: no joint pain and morning stiffness up to 1 h, arthrotic pain and morning stiffness up to 1 h, arthrotic pain and morning stiffness longer than 1 h.

3.3. The experiments

To evaluate our method for typical co-occurrences acquisition, we took the data set and the six groupings described above, the latter without the typical co-occurrences provided by the expert. We then applied our method to produce the typical co-occurrences. For each grouping, the typical co-occurrences produced by $\mathsf{HINT}_{\mathsf{TCO}}$ are listed in the column labeled '$\mathsf{HINT}_{\mathsf{TCO}}$' of Table 4. For example, $\mathsf{HINT}_{\mathsf{TCO}}$ suggests that the typical co-occurrences for grouping 1 should be: no joint pain and morning stiffness up to 1 h, arthrotic pain and morning stiffness up to 1 h, arthritic pain and morning stiffness up to 1 h.

The groupings with the new typical co-occurrences suggested by $\mathsf{HINT}_{\mathsf{TCO}}$ were then provided as background knowledge in addition to the 462 training examples (patient records). This background knowledge was used to introduce a new attribute for each grouping (as explained in Section 1). The 462 examples augmented with the six new attributes (thus having in total 22 attributes) were fed to the rule induction system CN2 [5] and to a nearest neighbor classifier [6,10,24]. The goal of this was to evaluate the usefulness of the new attributes and in this way the usefulness of the typical co-occurrences proposed by $\mathsf{HINT}_{\mathsf{TCO}}$.

Table 4
The six groupings and their typical co-occurrences

	Specialist	$\mathsf{HINT}_{\mathsf{TCO}}$
(1) Joint pain, morning stiffness		
No pain, $\leq$1 h	●	●
Arthrotic, $\leq$1 h	●	●
Arthritic, $\leq$1 h		●
No pain, $>$1 h		
Arthrotic, $>$1 h		
Arthrotic, $>$1 h	●	
f_{CN2}	2	1
f_{NN}	0.345	0.353
(2) Spinal pain, morning stiffness		
No pain, $\leq$1 h	●	●
Spondylotic, $\leq$1 h	●	●
Spondylitic, $\leq$1 h		●
No pain, $>$1 h		
Spondylotic, $>$1 h		
Spondylitic, $>$1 h	●	
f_{CN2}	3	3
f_{NN}	0.545	0.643
(3) Sex, other pain		
Male no		●
Male, muscles		●
Male, thorax	●	
Male, heels	●	
Male, other		●
Female, no		●
Female, other		●
Other 7 combinations		
f_{CN2}	1	4
f_{NN}	0.080	0.096
(4) Joint pain, spinal pain		
No pain, no pain	●	●
Arthrotic, no pain	●	●
Arthritic, no pain	●	●
No pain, spondylotic	●	
Arthrotic, spondylotic		●
Arthritic, spondylotic		
No pain, spondylitic	●	
Arthrotic, spondylitic		
Arthritic, spondylitic	●	
f_{CN2}	9	8
f_{NN}	0.908	0.743
(5) Joint pain, spinal pain, painful joints		
No pain, no pain, 0	●	●
No pain, no pain, $1 \leq$joints≤ 5		●
No pain, spondylotic, 0	●	●
No pain, spondylitic, 0	●	●
Arthrotic, no pain, $1 \leq$joints≤ 5	●	●

Table 4 (Continued)

	Specialist	HINT$_{TCO}$
Arthrotic, spondylotic, $1 \leq$joints≤ 5		●
Arthrotic, spondylotic, $5 <$joints≤ 30		●
Arthritic, no pain, $1 \leq$joints≤ 5	●	●
Arthritic, no pain, $5 <$joints≤ 30	●	●
Arthritic, spondylitic, $1 \leq$joints≤ 5	●	
Other 25 combinations		
f_{CN2}	7	9
f_{NN}	0.757	0.834
(6) Swollen joints, painful joints		
0, 0	●	●
0, $1 \leq$joints≤ 5	●	●
0, $5 <$joints≤ 30	●	
0, $30 <$		●
$1 \leq$joints≤ 10, 0	●	●
$1 \leq$joints≤ 10, $1 \leq$joints≤ 5	●	
$1 \leq$joints≤ 10, $5 <$joints≤ 30	●	●
$1 \leq$joints≤ 10, $30 <$		
$10 <$, 0		
$10 <$, $1 \leq$joints≤ 5		
$10 <$, $5 <$joints≤ 30		
$10 <$, $30 <$		
f_{CN}	1	1
f_{NN}	0.331	0.392

Two metrics were used to evaluate the usefulness of the new attributes. The number of occurrences of each grouping (i.e. the new attribute corresponding to that grouping) in the set of rules induced by CN2 is listed in the rows marked f_{CN2}. The higher this number, the more relevant the grouping. The mutual information between the grouping and the diagnostic class, calculated as a weight for nearest neighbor classification [24] is listed in the rows marked f_{NN}. The mutual information [22] between two random variables is defined as the reduction in uncertainty concerning the value of one variable that is obtained when the value of the other variable is known. The mutual information between an attribute and the class tells us how useful the attribute is for classification: if an attribute provides no information about the class, the mutual information will be zero. The two measures have been used in earlier experiments to assess the utility of background knowledge in machine learning [9,15].

3.4. The results

For groupings 1, 2, 5, and 6, the typical co-occurrences derived by HINT$_{TCO}$ correspond reasonably well to those proposed by the specialist for rheumatic diseases. For these groups, while using the same (groupings 1, 2, and 6) or slightly higher number of co-occurrences (grouping 5), two thirds or more of the

Table 5
Number of possible colors for columns of partition matrix of grouping 4

Joint pain, spinal pain	No. colors
No pain, no pain	1
Arthrotic, no pain	1
Arthritic, no pain	1
No pain, spondylotic	2
Arthrotic, spondylotic	1
Arthritic, spondylotic	3
No pain, spondylitic	2
Arthrotic, spondylitic	3
Arthritic, spondylitic	4

co-occurrences originally proposed by the specialist were discovered by $\mathsf{HINT}_{\mathsf{TCO}}$. This is different to grouping 4, where less than one half of the co-occurrences match and to grouping 3, where there are no matches.

In terms of the mutual information evaluation metrics f_{NN}, the co-occurrences derived by $\mathsf{HINT}_{\mathsf{TCO}}$ score higher for all groupings with the exception of grouping 4. A similar behavior is observed when the number of appearances in CN2 induced rules f_{CN2} is used as an evaluation metric. Here, $\mathsf{HINT}_{\mathsf{TCO}}$ scores equal or higher for all but the groupings 1 and 4.

Overall, compared to the co-occurrences proposed by the specialist, $\mathsf{HINT}_{\mathsf{TCO}}$ performed well for groupings 1, 2, 5, and 6. There are slight differences in the proposed co-occurrences, which, in turn, contribute to higher evaluation metric values. For grouping 3, there is a complete mismatch between the co-occurrences proposed by the specialist and those derived by $\mathsf{HINT}_{\mathsf{TCO}}$. The co-occurrences derived by $\mathsf{HINT}_{\mathsf{TCO}}$ score higher on both metrics (4 to 1 on f_{CN2}). However, the weights assigned by mutual information suggest that this grouping might be substantially less important for classification than the others (f_{NN} of 0.096 and 0.080).

It is grouping 4 where the of co-occurrences derived by $\mathsf{HINT}_{\mathsf{TCO}}$ seem to be less appropriate than those proposed by the specialist. However, note that for this grouping the specialist proposed six co-occurrences while $\mathsf{HINT}_{\mathsf{TCO}}$ discovered only four. Instead of using $\mathsf{HINT}_{\mathsf{TCO}}$ to derive only the typical co-occurrences for which the corresponding number of colors in the partition matrix is one, we can use this number as a measure of appropriateness for a certain combination of attribute values to be used as a typical co-occurrence. The lower the number of colors, the better the corresponding combination. For grouping 4, the number of possible colors for the columns in the partition matrix is shown in Table 5. It indicates that (no pain, spondylotic) and (no pain, spondylitic) are the next best candidates for typical co-occurrences. Interestingly, both are also proposed by the specialist. Their inclusion to the set of typical co-occurrences derived by $\mathsf{HINT}_{\mathsf{TCO}}$ makes this set very similar to that of the specialist, and also increases the mutual information weight from 0.743 to 0.887.

With the above extension, we can therefore conclude that $\mathsf{HINT}_{\mathsf{TCO}}$ discovered typical co-occurrences that were comparable to those proposed by the expert both in terms of similarity and usefulness as background knowledge for machine learning. This is important since $\mathsf{HINT}_{\mathsf{TCO}}$ is not meant to be a stand-alone tool for the unsupervised discovery of background knowledge, but should rather provide support to the expert by: (1) proposing a set of co-occurrences; and (2) weighting different combinations of attribute values to indicate how important it is that they are included in such a set. It would then be up to the expert to decide which of the proposed co-occurrences are meaningful and should be used.

As an overall evaluation of the typical co-occurrences suggested by $\mathsf{HINT}_{\mathsf{TCO}}$, let us consider the performance and size of the rules induced by CN2. The size was measured by the number of rules induced and the total number of rule conditions used. The performance measures used were the accuracy and information content (also called information score) [8,13]. The information score is a performance measure which is not biased by the prior class distribution: it takes into account the difficulty of the classification problem considered. It accounts for the possibility to achieve high accuracy easily in domains with a very likely majority class: classifying into the majority class all the time gives a zero information score.

To assess the performance, we used stratified 10-fold cross-validation [17]. This divides the data set into 10 sets of approximately equal size and class distribution. In each experiment, a single set is used for testing the classifier that has been developed from the remaining nine sets. The performance is subsequently assessed as an average of 10 experiments. Several experiments were performed that used the same training and testing data sets but obtained background knowledge differently and used:

1. no background knowledge,
2. typical co-occurrences for groupings as proposed by expert,
3. typical co-occurrences for the same groupings but derived by $\mathsf{HINT}_{\mathsf{TCO}}$ from the complete data set,
4. typical co-occurrences for the same groupings but derived by $\mathsf{HINT}_{\mathsf{TCO}}$ from training data sets.

The results (Table 6) indicate that both the accuracy and information content are higher when background knowledge is used. Best performance in terms of accuracy and information content is observed when typical co-occurrences are derived by $\mathsf{HINT}_{\mathsf{TCO}}$ from the complete data set. This, however, means that testing cases are also taken into account when deriving the typical co-occurences.

A realistic evaluation is obtained only if the typical co-occurrences are derived from the training examples alone. In this case, the performance is slightly lower, but comparable to the best performance. Using typical co-occurrences derived by $\mathsf{HINT}_{\mathsf{TCO}}$ from the training examples is clearly beneficial and yields comparable results as when using typical co-occurrences provided by a medical domain expert.

4. Computer-assisted selection of attribute groupings

In the experiments described above, $\mathsf{HINT}_{\mathsf{TCO}}$ assumed that the set of attributes for which to derive typical co-occurrences was given in advance. A possible extension

of this approach is to propose not only typical co-occurrences but also the set of attributes for which the background knowledge in the form of typical co-occurrences should be defined. The idea is straightforward and is illustrated with Algorithm 3. The algorithm examines all groupings of attributes from a candidate set (e.g. all pairs and triples) and for each grouping derives the set of typical co-occurrences. Each grouping is then assigned a weight, which estimates the usefulness of the grouping for classification. The groupings, sorted by decreasing weight, are then presented to the user who decides which of the proposed groupings are meaningful and are in his opinion suitable for use as background knowledge.

Algorithm 3: Derivation of groups of (two and three) attributes for which background knowledge in the form of typical co-occurrences might be useful for machine learning

---

Input: set of examples
Output: sorted list of attribute groupings with assigned weights

derive a decision table from the set of examples
for all the pairs (and triples) of attributes **do**
  derive the typical co-occurrences
  derive the corresponding weight
endfor
sort the groupings by descending weights and present them to the user

---

We have used this idea to obtain a list of sorted groupings of two attributes for the data set on early diagnosis of rheumatic diseases. The groupings were ranked according to the mutual information [22] between the attribute obtained from the grouping and the diagnostic class. While all five two-attribute groupings from Table 4 (originally proposed by the expert) ranked in the upper half of the sorted list of groupings, grouping 4 and grouping 2 were ranked within the best six groupings, which were:

1. 'Spinal pain' and 'Swollen joints'
2. 'Number of painful joints' and 'Spinal pain'
3. 'Spinal pain' and 'Skin manifestations'

Table 6
Accuracy and information content for rules induced by CN2

Background knowledge	Accuracy (%)	Inf. content (%)	No. rules	No. condition
None	43.1 ± 5.6	17.5 ± 3.2	27.7 ± 1.4	93.2 ± 4.4
TCO by expert	48.0 ± 3.3	24.8 ± 4.7	37.1 ± 3.3	113.8 ± 8.8
TCO by $\mathsf{HINT}_{\mathsf{TCO}}$, entire data set	48.7 ± 4.3	25.7 ± 3.5	36.1 ± 2.5	108.7 ± 8.8
TCO by $\mathsf{HINT}_{\mathsf{TCO}}$, training sets	46.8 ± 3.9	24.0 ± 3.3	$38.1 \pm 3\ 0$	115.9 ± 6.2

Means and standard deviations are given as estimated by 10-fold cross validation.

4. 'Joint pain' and 'Spinal pain' (grouping 4)
5. 'Spinal pain' and 'Therapy'
6. 'Spinal pain' and 'Morning stiffness' (grouping 2)

Note that all six highest ranked groupings include 'Spinal pain'. This may be contributed to by the high mutual information between the attribute itself and the class, which is also the highest among all nominal attributes used in the rheumatic diseases data set.

For an additional experiment, we have used 10-fold cross validation, and for each training set: (1) let $\mathsf{HINT}_{\mathsf{TCO}}$ choose six best groupings of two attributes; (2) derive the typical co-occurrences for these groupings; and (3) used CN2 to build a classifier whose performance was then assessed on the test set. The obtained accuracy and information score were $48.0 \pm 3.0\%$ and $26 \pm 3.8\%$, respectively. This result is comparable to the best performance in Table 6.

5. Conclusions

Background knowledge in the form of typical co-occurrences can have a positive effect on machine learning results in terms of the performance and the quality of induced rules from the point of view of comprehensibility. We have developed a method that proposes typical co-occurrences through functional decomposition of a given set of examples. In an earlier case study that we have re-considered in this paper, medical diagnosis background knowledge of this type has been completely specified by a medical expert. Our approach offers the possibility to automate the background knowledge acquisition process by proposing typical co-occurrences to the expert, who would then consider them in the light of his expert knowledge.

Experiments indicate that the use of typical co-occurrences identified by our method improves the performance of machine learning. The overall performance is comparable to the one obtained by using typical co-occurrences provided by a medical expert. While potentially useful attribute groupings can also be identified automatically, the involvement of the expert is crucial to obtain more comprehensible classification rules.

As further work, a more careful evaluation of the background knowledge acquired through using our method is needed. This should include an evaluation of the quality of induced rules from a medical point of view. Furthermore, experiments in other domains with an active involvement of a domain expert in both attribute grouping and typical co-occurrence selection should be conducted.

References

[1] Ashenhurst, RL. The decomposition of switching functions. Technical report, Bell Laboratories BL-1(11), 1952:541–602.
[2] Biermann AW, Fairfield J, Beres T. Signature table systems and learning. IEEE Trans. Syst. Man Cybern. 1982;12(5):635-648.

[4] Bohanec M, Zupan B, Bratko I, Cestnik B. A function decomposition method for development of hierarchical multi-attribute decision models. In: Proceedings of the Fourth Conference of the lnternational Society for Decision Support Systems (ISDSS-97). Switzerland: Lausanne, July 1997:503–514.
[5] Clark P, Boswell R. Rule induction with CN2: Some recent improvements. In Proc. 5th European Working Session on Learning. Berlin: Springer, 1991:151–163.
[6] Cover TM, Hart PE. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 1968;13:21-27.
[7] Curtis HA. A New Approach to the Design of Switching Functions. New York: Van Nostrand Reinhold, 1962.
[8] Džeroski S, Cestnik B, Petrovski I. Using the *m*-estimate in rule induction. J. Comput. Inf. Technol. 1993;1:37-46.
[9] Džeroski S, Lavrač N. Rule induction and instance-based learning applied in medical diagnosis. Technol. Health Care 1996;4:203-221.
[10] Fix E, Hodges JL. Discriminatory analysis. Nonparametric discrimination. Consistency properties. Technical Report 4, US Air Force School of Aviation Medicine. Randolph Field, Texas, 1957.
[11] Harmon P, Maus R, Morrissey W. Expert systems: Tools and Applications. New York: Wiley, 1988.
[12] Karalič A, Pirnat V. Machine learning in rheumatology. Sistemica 1990;1(2):113-123.
[13] Kononenko I, Bratko I. Information-based evaluation criterion for classifier's performance. Mach. Learn. 1991;6(1):67-80.
[14] Lavrač N, Džeroski S, Pirnat V, Križman V. Learning rules for early diagnosis of rheumatic diseases. In: Proc. 3rd Scandinavian Conference on Artificial Intelligence. Amsterdam: IOS Press, 1991:138–149.
[15] Lavrač N, Džeroski S, Pirnat V, Križman V. The utility of background knowledge in learning medical diagnostic rules. Appl. Artif. Intell. 1993;7:273-293.
[16] Michie D. Problem decomposition and the learning of skills. In: Lavrač N, Wrobel S, editors. Machine Learning: ECML-95. Notes in Artificial Intelligence 912. Berlin: Springer, 1995:17–31.
[17] Michie D, Spiegelhalter DJ, Taylor CC, editors. Machine learning, neural and statistical classification. Chichester: Ellis Horwood, 1994.
[18] Perkowski MI. Unified approach to functional decompositions of switching functions. Technical report, Warsaw University of Technology and Eindhoven University of Technology, 1995.
[19] Pirnat V, Kononenko I, Janc T, Bratko I. Medical analysis of automatically induced diagnostic rules. In: Proc. 2nd European Conference on Artificial Intelligence in Medicine. Berlin: Springer, 1989:24–36.
[20] Saaty TL. Multicriteria decision making: The analytic hierarchy process. RWS Publications, 1993.
[21] Samuel A. Some studies in machine learning using the game of checkers II: Recent progress. IBM J. Res. Develop. 1967;11:601-617.
[22] Shanon CE. A mathematical theory of communication. Bell. Syst. Tech. J. 1948;27:379-423.
[23] Shapiro AD, Niblett T. Automatic induction of classification rules for a chess endgame. In: Clarke MRB, editor. Advances in Computer Chess 3. Oxford: Pergamon, 1982:73–92.
[24] Wettschereck D. A study of distance-based machine learning algorithms. PhD thesis, Department of Computer Science, Oregon State University, Corvallis, OR, 1994.
[25] Zupan B, Bohanec M, Bratko I, Demšar J. Machine learning by function decomposition. In: Fisher JDH, editor. Proc. 14th International Conference on Machine Learning. San Mateo, CA: Morgan Kaufmann, 1997:421–429,