Acquiring and validating background knowledge for machine learning using function decomposition

Blaž Zupan and Sašo Džeroski

Department of Intelligent Systems, Jožef Stefan Institute
1000 Ljubljana, Slovenia (E-mail: Blaz.Zupan@ijs.si, Saso.Dzeroski@ijs.si)

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 to automate the process of acquiring background knowledge of this kind. 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.

1 Introduction

When applying 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*. While most machine learning approaches have only limited capabilities of taking into account such knowledge, inductive logic programming [12] systems can handle different types of background knowledge.

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

Typical co-occurrences are used in expert diagnosis. When asked for some additional knowledge about the difficult problem of early diagnosis of rheumatic diseases, a medical expert provided typical co-occurrences for several groupings of attributes. These were then used by the LINUS [12] system for inductive logic programming in the domain of early diagnosis of rheumatic diseases [14] from anamnestic data. In this domain, the task here is to diagnose into one of eight diagnostic classes, given sixteen anamnestic attributes. The difficulty of the diagnostic problem itself and noise in the data make this a very hard problem for machine learning approaches. A more detailed description of the domain can be found in Section 3.

The medical expert provided 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, LINUS introduces a new attribute which is considered in the learning process. For a particular patient record (example) this attribute has as 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 [3], or any attribute-value learning system can then 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" and the typical co-occurrences are: no spinal pain and morning stiffness up to 1 hour, spondylotic pain and morning stiffness up to 1 hour, spondylitic pain and morning stiffness longer than 1 hour. An example rule that uses this grouping and the second co-occurrence is given in Table 1. This rule was induced by LINUS using CN2 [14].

Table 1. A rule that makes use of a typical co-occurrence in the domain of early 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 hour'
THEN Diagnosis = Degenerative_spine_diseases
```

The background knowledge in the form of typical co-occurrences was shown to have positive effect on rule induction in several respects. First, rules induced in the presence of background knowledge performed better in terms of classification accuracy and information content [14]. Second, it substantially improves the quality of induced rules from a medical point of view as assessed by a medical expert [14]. Finally, it reduces the effects of noisy data on the process of rule induction and nearest neighbor classification [7].

The motivation for our work is based on the following line of reasoning: It is very desirable to have and use background knowledge in the form of typical co-occurrences in rule induction, as it can greatly improve performance. Typical co-occurrences are also a natural and useful human concept used by the medical expert. However, it is well-known that direct knowledge acquisition from experts is an arduous and error-prone process [9]. This paper therefore 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.

Before proceeding further, let us briefly mention related work. The domain of early diagnosis of rheumatic diseases has been first treated with a machine learning approach by Pirnat et al. [16]. Decision tree based approaches have been further applied to this domain by Karalič and Pirnat [10]. The use of background knowledge in this domain has been investigated by Lavrač et al. in combination with a decision tree approach [13] and in combination with a rule induction approach [14] and by Džeroski and Lavrač [7] in combination with nearest neighbor classification.

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 [5]. They have used function decomposition for the discovery of Boolean functions. Its potential use within machine learning was first observed by Samuel [17] and Biermann [2]. A recent report of Perkowski et al. [15] 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 and were developed by Zupan and Bohanec [20].

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 given in Section 3 and discussed in Section 4, where the typical co-occurrences provided by the expert and the ones generated by the proposed method are compared. The latter are quite similar to the former, but mostly have higher mutual information with the diagnostic class. Section 5 outlines some prospects for further fork and Section 6 concludes.

2 The method

This section formally and through an example introduces the method that, given a set of examples represented as attribute-value vectors with assigned classes, derives typical co-occurrences for a given set of attributes. The overall data-flow of the method is shown in Figure 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 attributes x_1, x_2, and x_3, and class y, and we want to decompose it to

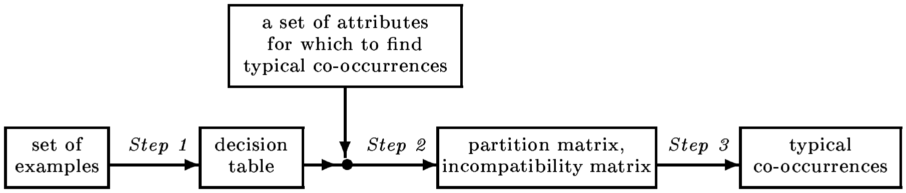

Fig. 1. The entities used and derived by the typical co-occurrence derivation method

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.

Table 2. A small 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

To derive G and H from F, we first need to represent a decision table with a *partition matrix* (Table 3). This uses all possible combinations of attribute values from the bound set as column labels and 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 of the function F.

Table 3. Partition matrix for the decision table from Table 2, 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

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 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 Figure 2.

For better comprehensibility, we interpret the column labels "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 lookup the annotated partition matrix for all the possible combinations of x_1

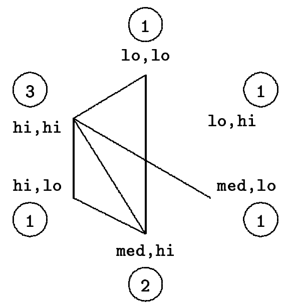

Fig. 2. Incompatibility graph for the partition matrix in Table 3

and c. The final result of the decomposition is represented as a hierarchy of two decision tables in Figure 3. If we further examine the discovered functions G and H we can see that $G \subset$ MAX and $H \subset$ MIN.

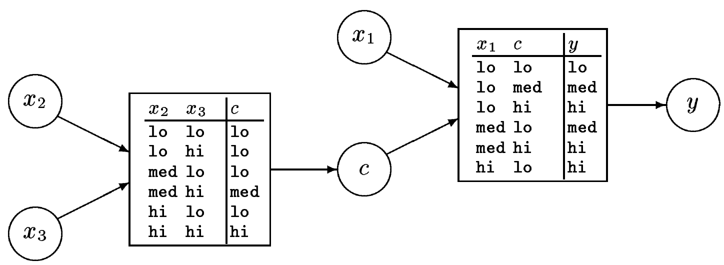

Fig. 3. The result of decomposing the decision table from Table 2

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 be assigned just a single color because of the incompatibilities with `(med,hi)` and `(hi,hi)` which are assigned different colors. While there exists just one distinct behavior 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 more about the behavior of the function F and is therefore more typical. Moreover, the columns that can be assigned only one color form a foundation for such color assignment 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 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 [20]. The typical co-occurrences derivation method then 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 about 15 nodes. Instead, we use the simple Color Influence Method of polynomial complexity [15]. The Color Influence Method sorts the nodes to color by their 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 $n \in IG$ 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 then used to augment the Color Influence Method to both color the incompatibility graph and discover typical co-occurrences (Algorithm 1).

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

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

Let us illustrate the use of Algorithm 1 on the incompatibility graph from Figure 2. The nodes sorted by decreasing connectivity are
`(hi,hi), (med,hi), (lo,lo), (hi,lo), (med,lo), (lo,hi)`
First, the node `(hi,hi)` is selected, determined to be typical (no other nodes have been colored yet), and assigned the color 1. Next, the node `(med,hi)` is considered. There are no colored nodes non-adjacent to it and so this node is typical. Since the adjacent node `(hi,hi)` has color 1, the color 2 is assigned to `(med,hi)`. Similarly, `(lo,lo)` is also typical and colored with 3 because the colors 1 and 2 have already been used for the adjacent nodes `(hi,hi)` and `(med,hi)`. Next, the 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 non-typical 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, the color 3 is then

arbitrarily chosen for `(med,lo)`. Similarly, the node `(lo,hi)` is found not to be typical and among three candidate colors the 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 such set. An alternative method that would search more exhaustively and possibly evaluate all different coloring of the incompatibility graph may be more complete and propose different sets of typical co-occurrences, but 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, the data sets from medical domains often include continuous attributes and may use several examples with the same attribute values but possibly 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.

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

while $E \neq \emptyset$ **do**
  find all $E' = \{e_k; e_k \in E\}$ 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 examples in E'
  add e_j with discretized continuous values and with class c to DT
  $E \leftarrow E \setminus E'$
endwhile

Algorithm 2: Derivation of a decision table from a set of 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 repeated until there are no more examples in E.

2.4 Implementation

The typical co-occurrences extraction method was implemented as $\mathsf{HINT}_{\mathsf{TCO}}$, an extension of the Hierarchy Induction Tool HINT [20] for learning concept hierarchies from examples by decision table decomposition. Both HINT and $\mathsf{HINT_{TCO}}$ run on a variety of UNIX platforms, including HP/UX, SunOS and IRIS.

3 Extracting and validating background knowledge in early diagnosis of rheumatic diseases

3.1 The domain

The data used originate from the University Medical Center in Ljubljana [16] and comprise records on 462 patients. The multitude of over 200 different diagnoses have been grouped into three, six, eight or twelve diagnostic classes. Our study uses eight diagnostic classes: degenerative spine diseases, degenerative joint diseases, inflammatory spine diseases, other inflammatory diseases, extraarticular rheumatism, crystal-induced synovitis, non-specific rheumatic manifestations, and non-rheumatic diseases.

For each patient, sixteen 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 4.1 we can see that the attribute "Duration of morning stiffness" has been discretized into two intervals: up to 1 hour and longer than 1 hour.

3.2 The background knowledge

In an earlier study [13], a specialist for rheumatic diseases has 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 dot in the row marked "specialist" and the column marked X means that tuple X is a typical co-occurrence for the corresponding Grouping. For example, Table 4.1 specifies Grouping 1, which relates the attributes "Joint pain" and "Duration of morning stiffness", with typical co-occurrences suggested by $\mathsf{HINT_{TCO}}$: no joint pain and morning stiffness up to 1 hour, arthrotic pain and morning stiffness up to 1 hour, arthritic pain and morning stiffness up to 1 hour.

3.3 The experiments

To evaluate our method for typical co-occurrences acquisition, we took the dataset 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 $\mathsf{HINT_{TCO}}$ are listed in the row labeled "$\mathsf{HINT_{TCO}}$" of the corresponding table. 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 hour, arthrotic pain and morning stiffness up to 1 hour, arthritic pain and morning stiffness up to 1 hour.

The groupings with the new typical co-occurrences suggested by $\mathsf{HINT}_{\mathsf{TCO}}$ are then provided as background knowledge to LINUS [12] in addition to the 462 training examples (patient records). LINUS then introduces a new attribute for each grouping (as explained in the introduction). The 462 examples augmented with the six new attributes (thus having in total 22 attributes) are then fed to the rule induction system CN2 [3] and to a nearest neighbor classifier [19, 8, 4]. 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 $\mathsf{HINT}_{\mathsf{TCO}}$.

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 column marked f_{CN2}. The mutual information between the grouping and the diagnostic class, calculated as a weight for nearest neighbor classification [19] is listed in the column marked f_{NN}. The mutual information [18] between an attribute and the class tells us how useful the attribute is for classification. The two measures have been used in earlier experiments to assess the utility of background knowledge in machine learning [14, 7].

4 Discussion

For groupings 1, 2, 5, and 6, the typical co-occurrences derived by $\mathsf{HINT}_{\mathsf{TCO}}$ 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 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 but the grouping 4. A similar behavior is observed when the number of appearances in CN2 induced rules f_{CN2} is used as an evaluation metrics. There, $\mathsf{HINT}_{\mathsf{TCO}}$ 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 values of the evaluation metrics. 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_{CN2} 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

Table 4. The six groupings and their typical co-occurrences

Joint pain, Morning stiffness	specialist	HINT$_{TCO}$
No pain, $\leq$ 1 hour	•	•
Arthrotic, $\leq$ 1 hour	•	•
Arthritic, $\leq$ 1 hour		•
No pain, $>$ 1 hour		
Arthrotic, $>$ 1 hour		
Arthritic, $>$ 1 hour	•	
f_{CN2}	2	1
f_{NN}	0.345	0.353

1)

Spinal pain, Morning stiffness	specialist	HINT$_{TCO}$
No pain, $\leq$ 1 hour	•	•
Spondylotic, $\leq$ 1 hour	•	•
Spondylitic, $\leq$ 1 hour		•
No pain, $>$ 1 hour		
Spondylotic, $>$ 1 hour		
Spondylitic, $>$ 1 hour	•	
f_{CN2}	3	3
f_{NN}	0.545	0.643

2)

Sex, Other pain	specialist	HINT$_{TCO}$
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

3)

Joint pain, Spinal pain	specialist	HINT$_{TCO}$
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

4)

Joint pain, Spinal pain, Painful joints	specialist	HINT$_{TCO}$
No pain, No Pain, 0	•	•
No pain, No Pain, 1$\leq$joints$\leq$5		•
No pain, Spondylotic, 0	•	•
No pain, Spondylitic, 0	•	•
Arthrotic, No pain, 1$\leq$joints$\leq$5	•	•
Arthrotic, No pain, 5$<$joints$\leq$30	•	•
Arthrotic, Spondylotic, 1$\leq$joints$\leq$5		•
Arthrotic, Spondylotic, 5$<$joints$\leq$30		•
Arthritic, No pain, 1$\leq$joints$\leq$5	•	•
Arthritic, No pain, 5$<$joints$\leq$30	•	•
Arthritic, Spondylitic, 1$\leq$joints$\leq$5	•	
other 25 combinations		
f_{CN2}	7	9
f_{NN}	0.757	0.834

5)

Swollen joints, Painful joints	specialist	HINT$_{TCO}$
0, 0	•	•
0, 1$\leq$joints$\leq$5	•	•
0, 5$<$joints$\leq$30	•	
0, 30$<$		•
1$\leq$joints$\leq$10, 0	•	•
1$\leq$joints$\leq$10, 1$\leq$joints$\leq$5	•	
1$\leq$joints$\leq$10, 5$<$joints$\leq$30	•	•
1$\leq$joints$\leq$10, 30$<$		
10$<$, 0		
10$<$ 1$\leq$joints$\leq$5,		
10$<$, 5$<$joints$\leq$30		
10$<$, 30$<$		
f_{CN2}	1	1
f_{NN}	0.331	0.392

6)

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-occurrences. 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.

Table 5. Number of possible colors for columns of partition matrix of Grouping 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
# colors	1	1	1	2	1	3	2	3	4

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 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 from the dataset generated by LINUS. The accuracy and information content [11, 6] (as measured on the training set) of the rules induced (using the significance test in CN2) are 56.5% and 31%, respectively. For comparison, those obtained with the expert-proposed co-occurrences are 52.4% and 30%, respectively. The co-occurrences proposed by $\mathsf{HINT}_{\mathsf{TCO}}$ yield 35 rules with 106 conditions, while the expert-proposed ones yield 38 rules with 120 conditions. CN2 without background knowledge performs worse than with either of the two sets of co-occurrences: the accuracy and information content are 51.7% and 22%, while the rule set contains 30 rules and 102 conditions.

5 Further work

Currently $\mathsf{HINT}_{\mathsf{TCO}}$ assumes the set of attributes for which to derive typical co-occurrences are given in advance. We envision an extension of this approach to propose not only co-occurrences but also the set of attributes for which the background knowledge in the form of co-occurrences should be defined. The idea is straightforward and is illustrated with Algorithm 5. The implementation would require the integration of $\mathsf{HINT}_{\mathsf{TCO}}$ with the machine learning tools that evaluate and use the groupings.

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

---

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

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. An evaluation of the performance in terms of classification accuracy on unseen cases is also desirable, but requires a slightly more complicated experimental setup: typical co-occurrences would have to be determined for each partition of the dataset into training and testing cases. An alternative is to have a medical expert assess the co-occurrences suggested by $\mathsf{HINT}_{\mathsf{TCO}}$, which is the most desirable option.

6 Conclusion

Background knowledge in the form of typical co-occurrences has positive effect on machine learning results in terms of performance and the quality of induced rules from a medical point of view. We have developed a method that proposes typical co-occurrences through functional decomposition of a given set of examples. While medical diagnosis background knowledge of this type has been previously 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.

References

1. Ashenhurst, R.L. (1952). The decomposition of switching functions. Technical report, Bell Laboratories BL-1(11): 541–602.

2. Biermann, A.W., Fairfield, J., and Beres, T. (1982). Signature table systems and learning. *IEEE Trans. Syst. Man Cybern.*, 12(5): 635–648.
3. Clark, P., and Boswell, R. (1991). Rule induction with CN2: Some recent improvements. In *Proc. Fifth European Working Session on Learning*, pages 151–163. Springer, Berlin.
4. Cover, T.M., and Hart, P.E. (1968). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13: 21–27.
5. Curtis, H.A. (1962). *A New Approach to the Design of Switching Functions.* Van Nostrand, Princeton, N.J.
6. Džeroski, S., Cestnik, B., and Petrovski, I. (1993). Using the m-estimate in rule induction. *Journal of Computing and Information Technology*, 1: 37–46.
7. Džeroski, S., and Lavrač, N. (1996). Rule induction and instance-based learning applied in medical diagnosis. *Technology and Health Care.*
8. Fix, E., and Hodges, J.L. (1957). Discriminatory analysis. Nonparametric discrimination. Consistency properties. Technical Report 4, US Air Force School of Aviation Medicine. Randolph Field, TX.
9. Harmon, P., Maus, R., and Morrissey, W. (1988). *Expert systems: Tools & Applications.* John Wiley, New York.
10. Karalič, A., and Pirnat, V. (1990). Machine learning in rheumatology. *Sistemica* 1(2): 113–123.
11. Kononenko, I., and Bratko, I. (1991). Information-based evaluation criterion for classifier's performance. *Machine Learning*, 6(1): 67–80.
12. Lavrač, N., and Džeroski, S. (1994). *Inductive Logic Programming: Techniques and Applications.* Ellis Horwood, Chichester.
13. Lavrač, N., Džeroski, S., Pirnat, V., and Križman, V. (1991). Learning rules for early diagnosis of rheumatic diseases. In *Proc. Third Scandinavian Conference on Artificial Intelligence*, pages 138–149. IOS Press, Amsterdam.
14. Lavrač, N., Džeroski, S., Pirnat, V., and Križman, V. (1993). The utility of background knowledge in learning medical diagnostic rules. *Applied Artificial Intelligence*, 7:273–293.
15. Perkowski, M.A., and Grygiel, S. (1995). A survey of literature on function decomposition. Techical report, Dept. of Electrical Engineering, Portland State University.
16. Pirnat, V., Kononenko, I., Janc, T., and Bratko, I. (1989). Medical analysis of automatically induced rules. In *Proc. 2nd European Conference on Artificial Intelligence in Medicine*, pages 24–36. Springer, Berlin.
17. Samuel, A. (1967). Some studies in machine learning using the game of checkers II: Recent progress. *IBM J. Res. Develop.*, 11:601–617.
18. Shannon, C.E. (1948). A mathematical theory of communication. *Bell. Syst. Techn. J.*, 27: 379–423.
19. Wettschereck, D. (1994). A study of distance-based machine learning algorithms. PhD Thesis, Department of Computer Science, Oregon State University, Corvallis, OR.
20. Zupan, B., and Bohanec, M. (1996). Learning concept hierarchies from examples by function decomposition. Technical report, IJSDP–7455, J. Stefan Institute, Ljubljana. URL `ftp://ftp-e8.ijs.si/pub/reports/IJSDP-7455.ps`.

This article was processed using the LaTeX macro package with LLNCS style