of node i from top of the graph to the current position of the node. The variation measure $v(G_j)$ for each graph is:

$$v(G_j) = \begin{cases} v_1(G_j) & : \quad h(G_j) = 1 \\ v_2(G_j) & : \quad h(G_j) > 1 \end{cases} \qquad (1)$$

where $v(G_j) = \{v_1(G_j)|v_2(G_j)\}$, and $v_1(G_j)$ and $v_2(G_j)$ are defined as:

$$\begin{aligned} v_1(G_j) &= \sum_{i=0}^{n} \frac{c(G_j)}{C(G_j)} . \frac{a(G_{ji})}{A(G_{ji})} . \frac{1}{n+1} \\ v_2(G_j) &= \sum_{i=0}^{n} \frac{c(G_j)}{C(G_j)} . \frac{a(G_{ji})}{A(G_{ji})} . \frac{h(G_j) - l(G_{ji})}{h(G_j) * n} \end{aligned}$$

The total dissimilarity measure $d(G)$ between G_{t-1} and G_t is then defined as:

$$d(G) = v(G_{t-1}) + v(G_t); 0 \leq d(G) \leq 1 \qquad (2)$$

Query Refinement using Unexpected Rules

This section describes the second technique to assist the consultant in uncovering new or unexpected knowledge. This technique allows the consultant to validate the quality of the rules and to ensure the data in the database is consistent. In addition, it makes the query process dynamic in the sense that the questions used in the query as well as the order in which they are presented changes to reflect the knowledge being accumulated over time. The technique uses a distance matrix on the attribute groupings obtained from the domain expert for measuring the unexpectedness of the discovered knowledge, and an association rule mining algorithm for generating rules that are hidden in the dataset and might not be included in the classification rules obtained by C4.5. However, association rule mining usually generates a large set of rules that are obvious to the domain experts. Therefore, it is important to eliminate these rules, and only display those that are unexpected. The rules are ranked from the most unexpected to the least unexpected.

Grouping of Similar Attributes

We have explored two different approaches for determining unexpected and commonsense rules.

1. Domain Expert Grouping: This part of the experiment uses the "Dermatology" dataset. The consultant is required to manually group the attributes in a hierarchy according to their type. Figure 4 shows how the groupings are structured as a tree. Leaf nodes represent all the possible attributes. The dissimilarity between each pair of attributes is determined by the distance between the two leaf nodes. We calculate the unexpectedness of a rule by taking the maximum dissimilarity between any pair of attributes.

   The distance measure between the attributes in the leaf nodes is defined as:

$$distance(i,j) = d_i + d_j - 2d_{p(i,j)} \qquad (3)$$

   where i and j are nodes representing antecedents in the association rules; d_i and d_j are the depth of nodes i and j, respectively; and $d_{p(i,j)}$ is the depth of the lowest common parent node of nodes i and j from root.

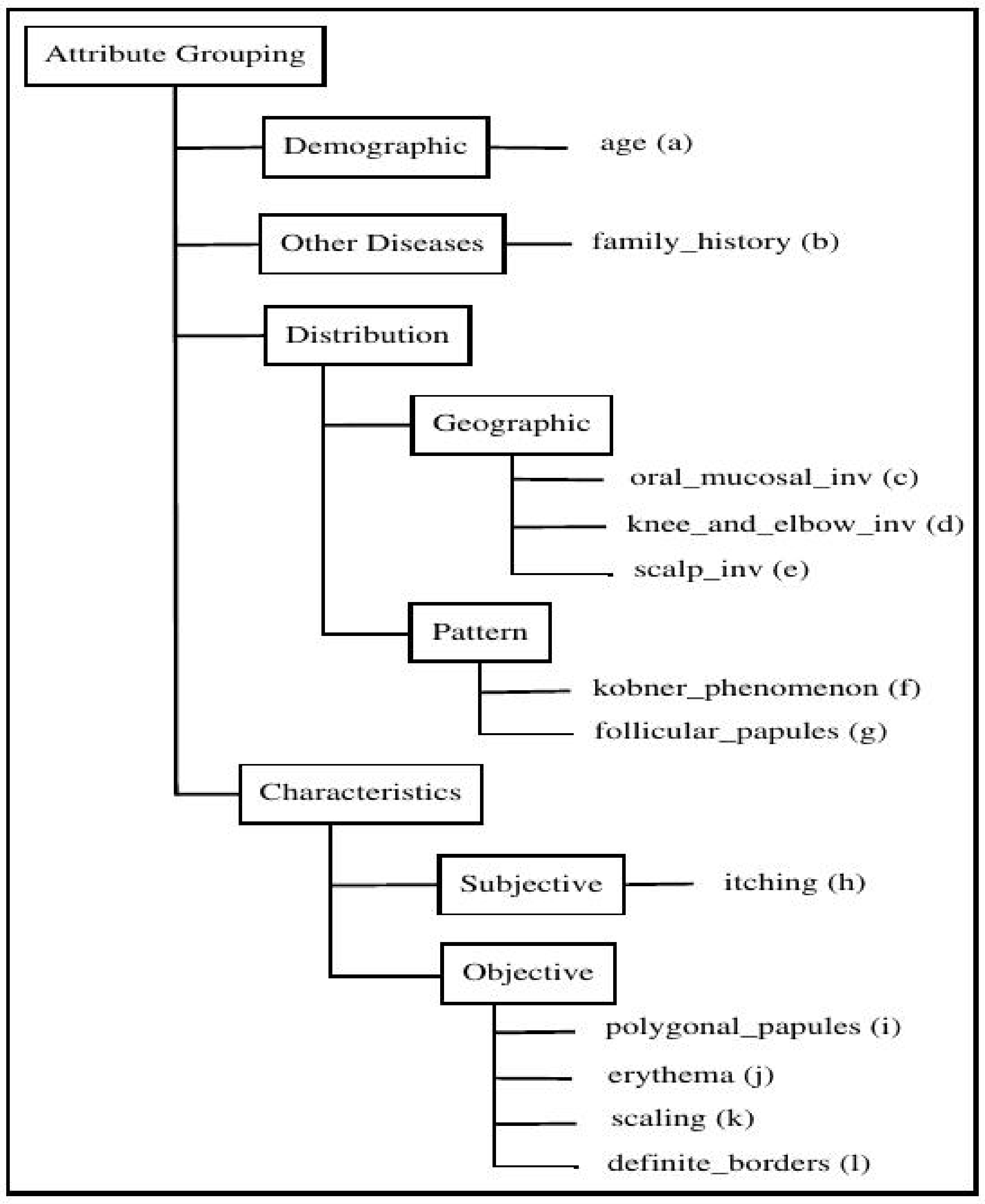

Figure 4: Hierarchical attribute grouping by the consultant.

   We apply Equation 3 to the groupings in Figure 4 to derive the distance matrix shown in Table 1. As can be seen from the table, the further apart the leaf nodes the greater the distance between them.

Table 1: Distance matrix derived from Figure 4.

	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

2. Decision Tree Classification Grouping: Decision trees only use features they need to unambiguously classify cases. However, there are many other useful features that exist in the cases that have not been extracted by the decision tree classifier. Therefore, we need to use an association rule mining technique to provide a full description of the diseases to assist the query process. This automatic approach dynamically generates a set of commonsense and unexpected classification-based association rules. To identify the commonsense/unexpected rules, the antecedence of the association rules are com-