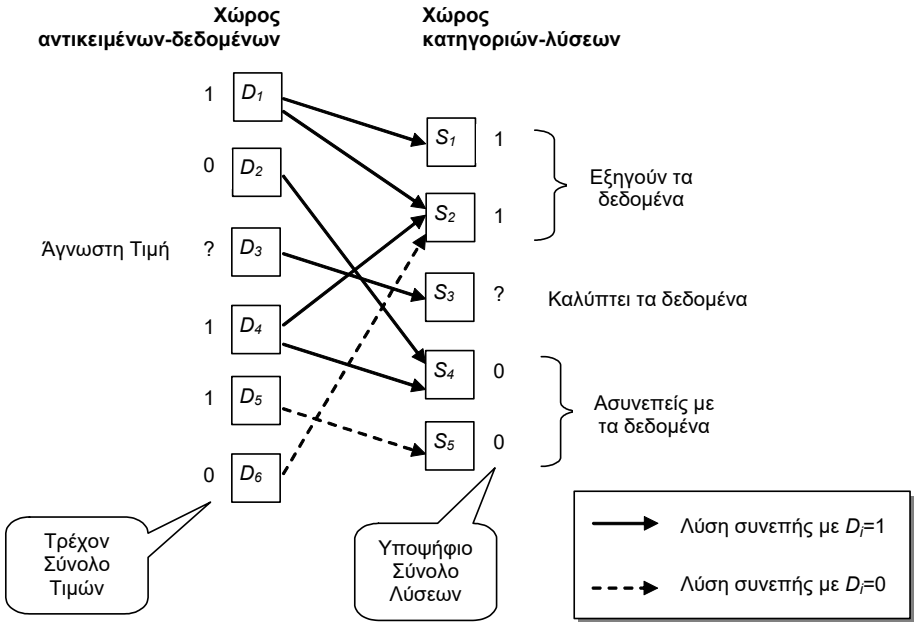
λύσης, όπου το 1 σημαίνει ότι η λύση είναι συνεπής, το 0 ότι είναι ασυνεπής και το ? ότι δεν υπάρχουν αρκετά στοιχεία για να αποφασιστεί. Στο Σχήμα 24.3, οι S_1 και S_2 ταιριάζουν με (ή εξηγούν) τα δεδομένα, η S_3 είναι συνεπής ή καλύπτει ορισμένα δεδομένα αλλά δεν τα εξηγεί και οι S_4 και S_5 είναι ασυνεπείς.

Ένα πλεονέκτημα αυτού του τρόπου αναπαράστασης είναι ότι για τον προσδιορισμό της συνέπειας μιας λύσης απαιτείται μόνο η διάδοση των τιμών από τα δεδομένα στις υποψήφιες λύσεις. Στις συνεχείς γραμμές οι τιμές διαδίδονται όπως είναι. Στις διακεκομμένες γραμμές οι τιμές αναστρέφονται, δηλαδή το 1 γίνεται 0 και αντίστροφα. Το "?" διαδίδεται πάντα ως έχει. Η κατάσταση μιας λύσης S_j προσδιορίζεται ως εξής:

- Αν όλες οι τιμές που διαδίδονται στο S_j είναι 1, τότε το S_j ταιριάζει ή εξηγεί τα δεδομένα.
- Αν κάποια τιμή που διαδίδεται στο S_j είναι 0 τότε το S_j είναι ασυνεπές με τα δεδομένα και απορρίπτεται.
- Αν όλες οι τιμές που διαδίδονται στο S_j είναι 1 και ? τότε το S_j είναι συνεπές ή καλύπτει τα δεδομένα.



Σχήμα 24.3: Παράδειγμα συζευκτικού μοντέλου κατηγοριοποίησης.

Υπάρχει πιθανότητα κάποια διανύσματα τιμών να μην ταιριάζουν με καμία κατηγορία ή να ταιριάζουν με περισσότερες από μια κατηγορίες. Στη δεύτερη περίπτωση ορισμένα συστήματα χρησιμοποιούν επιπρόσθετα κριτήρια για να προσδιορίσουν την κατηγορία. Το σύνολο των δεδομένων πιθανώς χρειάζεται να επεκταθεί από το σύνολο {0, 1} σε ένα άλλο σύνολο, πιο κατάλληλο για την εφαρμογή που προορίζεται. Επίσης ο συνδυασμός των συνθηκών πιθανώς να χρειάζεται να αλλάξει. Στο *συζευκτικό* μια κατηγορία ταιριάζει αν *όλες* οι συνθήκες ικανοποιούνται (σχέση AND*)*. Υπάρχουν και άλλοι τρόποι συνδυασμού όπως OR, XOR ή συνδυασμός λογικών τελεστών. Μια άλλη τροποποίηση του μοντέλου είναι η προσθήκη πιθανοτήτων και αβεβαιότητας στην εύρεση λύσεων (κατηγοριών).

24.3 Μέθοδοι Κατηγοριοποίησης

Ο βασικός στόχος των μεθόδων κατηγοριοποίησης είναι ο αποκλεισμός των εναλλακτικών μονοπατιών στο χώρο αναζήτησης που αποτελείται από υποψήφιες κατηγορίες των δεδομένων εισόδου. Στη συνέχεια, παρουσιάζονται τέσσερις μέθοδοι κατηγοριοποίησης που έχουν χρησιμοποιηθεί σε διάφορα συστήματα. Για συντομία, αποκαλούνται με τα ονόματα K1 έως K4.

24.3.1 Παραγωγή και Δοκιμή

Πρόκειται για την πιο απλή μέθοδο (K1) η οποία βασίζεται στην τεχνική της εξαντλητικής *παραγωγής και δοκιμής* (*exhaustive generate and test*). Με τη μέθοδο αυτή δοκιμάζονται όλες οι πιθανές κατηγορίες-λύσεις και ελέγχεται αν κάποια λύση ταιριάζει με τα δεδομένα. Ο αλγόριθμος της μεθόδου K1 είναι ο ακόλουθος:

```
1. Θέσε την κενή λίστα ως λίστα των κατηγοριών-λύσεων
2. Πάρε τα δεδομένα εισόδου και γενίκευσέ τα
3. Για κάθε υποψήφια γενική κατηγορία
   i.   Έλεγξε αν τα δεδομένα εισόδου ανήκουν στην υποψήφια γενική κατηγορία
   ii.  Εάν ναι, πρόσθεσε την υποψήφια γενική κατηγορία στη λίστα των
        κατηγοριών-λύσεων
4. Ανέφερε τις λύσεις από τη λίστα των κατηγοριών-λύσεων
```

Οι προϋποθέσεις για τη χρήση ενός τέτοιου αλγόριθμου είναι οι εξής:

- Το σύνολο των κατηγοριών (λύσεων) είναι αρκετά μικρό έτσι ώστε η εξαντλητική σύγκριση να είναι πρακτικά εφικτή.
- Όλα τα απαραίτητα δεδομένα μπορεί να αποκτηθούν στην αρχή της διαδικασίας, έτσι ώστε να θεωρηθεί ότι η απουσία κάποιων δεδομένων σημαίνει ότι αυτά δεν υφίστανται.

Ουσιαστικά, ελάχιστα συστήματα χρησιμοποιούν αυτήν τη μέθοδο, η οποία παρατίθεται ως βάση για την παρουσίαση των υπολοίπων μεθόδων.

24.3.2 Από τα Δεδομένα σε Πιθανές Λύσεις

Η μέθοδος αυτή (K2) μειώνει τον υπολογιστικό χρόνο, σε σχέση με τη μέθοδο K1, όταν υπάρχει μεγάλος αριθμός λύσεων, αφού επικεντρώνεται μόνο σε εκείνες που *πιθανώς* ικανοποιούν τα δεδομένα. Επιπλέον, ασχολείται με κάθε υποψήφια λύση μόνο μια φορά επειδή καταγράφει ποιες λύσεις έχει ήδη δοκιμάσει. Οι προϋποθέσεις χρήσης της K1 ισχύουν και σε αυτήν την περίπτωση. Η μέθοδος χρησιμοποιεί τις ακόλουθες ειδικές διαδικασίες:

- *Διαδικασία γενίκευσης δεδομένων* (*data abstractor*), η οποία χρησιμοποιεί τις τεχνικές γενίκευσης που αναφέρθηκαν και είναι από μόνη της ένα μικρό σύστημα κατηγοριοποίησης. Για κάθε σύνολο δεδομένων μπορεί να υπάρχουν περισσότερες από μία δυνατότητες γενίκευσης.
- *Διαδικασία ανάκλησης υποψήφιων λύσεων* (*candidate retriever*), οι οποίες μπορούν εν δυνάμει να *εξηγήσουν* ένα δεδομένο. Σύμφωνα με το συζευκτικό μοντέλο

κατηγοριοποίησης αυτή η διαδικασία μπορεί να υλοποιηθεί επιστρέφοντας όλες τις λύσεις που είναι *συνεπείς* με τα δεδομένα ή, εναλλακτικά, όλες τις λύσεις που *καλύπτουν* τα δεδομένα. Προφανώς εδώ μπορεί να υπάρξουν και ευρετικές διαδικασίες ανάκλησης λύσεων οι οποίες να ακολουθούν άλλο μοντέλο κατηγοριοποίησης. Όποια διαδικασία και αν χρησιμοποιηθεί δεν θα πρέπει να επιστρέφει πολύ λίγες υποψήφιες λύσεις, γιατί τότε θα χαθούν κάποιες από τις πραγματικές λύσεις, ούτε πάρα πολλές, γιατί τότε δε θα υπάρχει βελτίωση στην απόδοση σε σχέση με τη μέθοδο K1.

- *Διαδικασία ελέγχου λύσεων (solution tester)*, η οποία βαθμολογεί τις υποψήφιες κατηγορίες-λύσεις βασιζόμενος όχι σε πρόσθετα δεδομένα, αλλά σε άλλα κριτήρια, όπως για παράδειγμα προϋπάρχουσες πιθανότητες, και επιλέγει στο τέλος ένα υποσύνολο αυτών.

Η χρήση όλων των παραπάνω φαίνεται στον ακόλουθο αλγόριθμο:

```
1. Θέσε την κενή λίστα ως λίστα των κατηγοριών-λύσεων.
2. Πάρε τα δεδομένα εισόδου .
3. Γενίκευσε τα δεδομένα με τη διαδικασία data_abstractor.
4. Βρες τις υποψήφιες γενικές κατηγορίες με τη διαδικασία
   candidate_retriever.
5. Για κάθε υποψήφια γενική κατηγορία:
   i.    Έλεγξε αν τα δεδομένα εισόδου ανήκουν στην υποψήφια γενική
         κατηγορία με τη διαδικασία solution_tester.
   ii.   Εάν ναι, πρόσθεσε την υποψήφια γενική κατηγορία στη λίστα των
         κατηγοριών-λύσεων.
6. Ανέφερε τις λύσεις από τη λίστα των κατηγοριών-λύσεων.
```

24.3.3 Ιεραρχική Κατηγοριοποίηση Οδηγούμενη από τις Λύσεις

Η μέθοδος αυτή (K3) αποτελεί μια ιεραρχική τεχνική παραγωγής και δοκιμής, η οποία υποθέτει ότι οι πιθανές κατηγορίες-λύσεις είναι ιεραρχικά διατεταγμένες και διασχίζει την ιεραρχία από πάνω προς τα κάτω και πρώτα σε πλάτος. Οι τελικές κατηγορίες-λύσεις βρίσκονται στα φύλλα του δένδρου, ενώ οι λύσεις των ενδιάμεσων επιπέδων είναι μερικές ή γενικές. Σε κάθε επίπεδο, η μέθοδος συγκρίνει τις υποψήφιες λύσεις με τα δεδομένα και απορρίπτει κλαδιά του δένδρου. Επίσης, μπορεί να ζητήσει πρόσθετα δεδομένα για να επιτύχει λεπτομερέστερη διάκριση μεταξύ των υποψήφιων λύσεων. Στη συνέχεια, η μέθοδος προχωρά στον έλεγχο λύσεων σε κατώτερο επίπεδο του δένδρου, ακολουθώντας μόνο τα κλαδιά που δεν έχουν απορριφθεί. Αν μια κατηγορία απορριφθεί, είναι προφανές ότι απορρίπτονται και όλες οι υποκατηγορίες που περιλαμβάνει.

Ο αλγόριθμος της μεθόδου K3 είναι ο ακόλουθος:

```
1. Θέσε την κενή λίστα ως λίστα των κατηγοριών-λύσεων.
2. Για κάθε επίπεδο της ιεραρχίας των κατηγοριών-λύσεων, επανέλαβε τα
   ακόλουθα:
```