

Chapter 2

Machine Learning for Early DRG Classification

In this chapter, a literature review of machine learning methods is provided with a special focus on attribute selection and classification methods successfully employed in health care. Similarities and differences between the machine learning methods addressed in this dissertation and the approaches available from the literature are highlighted. Afterwards, techniques for selecting relevant and non-redundant attributes for early DRG classification are presented. Finally, different classification techniques are described in detail.

2.1 Machine Learning for Health Care: A Literature Review

Attribute selection and classification are central methods of machine learning. Rather than employing large attribute sets for predicting or classifying a variable, it can be more efficient to select attributes in a first stage and afterwards to employ the selected attributes for prediction or classification. Textbook references for attribute selection and classification techniques are, among others, Bishop [22], Mackay [131] and Witten and Frank [233]. When considering attribute selection, Yu and Liu [239] divide the attribute selection process into three parts: Searching for irrelevant, weakly relevant and strongly relevant features, respectively. First, irrelevant features are not informative with respect to the class and can safely be ignored. Second, the set of weakly relevant features comprises redundant and non-redundant features. Third, strongly relevant features are always necessary for determining an optimal subset of features since removing a strongly relevant feature would necessarily affect the original conditional class distribution. The optimal subset of features is therefore the use of strongly relevant features and features that are weakly relevant but non-redundant.

After selecting an optimal set of features, a variable is predicted. The prediction process is denoted as a classification problem when the variable consists of different classes or categories. In the simplest case, the class variable is binary. However, it can consist of multiple features as well, as we will see in the case of early DRG

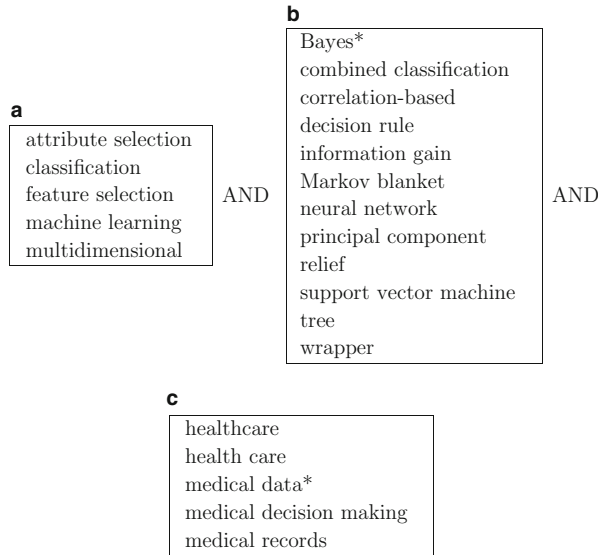


Fig. 2.1 Three-level structure of the search query including relevant tasks (a), machine learning methods (b) and field of research (c)

classification. In contrast, when predicting numeric or continuous attributes, those problems can be denoted as regression problems. In what follows, the focus is on classification problems.

2.1.1 Selection Criteria and Search for Relevant Literature

Attribute selection and classification methods are used in many health care applications and the following procedure describes a reproducible search for retrieving attribute selection and classification literature in the following fields: Health care, industrial engineering, medical informatics as well as operations research and the management sciences (OR/MS). English journal articles from the 2011 ISI journal citation report (JCR) science and social sciences edition (see [2]) were searched. Because of the aim to providing an interdisciplinary review, the subject categories anesthesiology, health policy and services, industrial engineering, medical informatics as well as OR/MS were selected within the JCR, yielding a total number of 206 relevant journals. Scopus (see [1]) was chosen as the search engine to retrieve the publications. The structure of the search query is presented in Fig. 2.1. The search string is divided hierarchically into (i) machine learning tasks (e.g., the selection of relevant attributes), (ii) machine learning methods (e.g. Markov blanket attribute selection) and (iii) field of research (e.g. health care). The asterisk (*) denotes a so-called “wildcard” which replaces character combinations of different length.

Table 2.1 Number of hits, irrelevant and remaining articles, categorized according to publication year

	1965–1997	1998–2002	2003–2007	2008–present	Total
#hits	32	55	87	154	328
#irrelevant articles	17	7	19	30	73
#remaining articles	15	48	68	124	255

For example, searching for “Bayes*” can result in, e.g., “Bayesian networks”. Inside the boxes, the search strings are connected by an “OR” such that, e.g., articles that consider multiple machine learning methods can be retrieved. Moreover, the “ANDs” between the boxes ensure that the search space contains at least one keyword of each box. Note that the search does not exclude any strings by using “AND NOT”, because false negative exclusions from the results should be avoided. However, exclusion criteria and the exclusion of irrelevant articles is described in the following section.

2.1.2 Classification of Relevant Literature

Table 2.1 presents summary statistics about the original 328 hits, excluded articles and the remaining 255 papers retrieved until May 22nd, 2013. An article is considered irrelevant if it focuses for example only on regression problems instead of classification problems (see, e.g., Bertsimas et al. [21]).

The table reveals that machine learning problems with a focus on health care are becoming increasingly popular in the literature and that in the last 5 years, more articles have been published than in any other of the previous time intervals. A forward search was performed, which means that more recent articles that cite the articles as given in Table 2.1 were included into the set of relevant papers. In addition, relevant contributions were discovered employing a backward search. This means that older articles that are cited within the references of the articles as given in Table 2.1 were reviewed and, if relevant, added to the set of relevant papers. In doing so, articles such as Hsieh et al. [88] and Kononenko et al. [107] were discovered using a forward and backward search, respectively. Within the set of relevant articles published since 1965, some literature reviews are available which are categorized in Table 2.2. The table shows that the set of journal articles contains nine literature reviews. Four of them focus on classification methods of which Harper [86] is the most recent. More specifically, Dreiseitl and Ohno-Machado [57] and Podgorelec et al. [166] limit their reviews to the methods neural networks and decision trees, respectively. However, the aim of this dissertations’s review is, to provide an overview of different standard attribute ranking and selection as well as classification techniques. Another drawback of the existing literature reviews is that, for example, Scotch et al. [198] focus exclusively on biomedical

Table 2.2 Literature reviews on classification methods, decision support systems and statistical analyses in health care

	Publication year	Classification methods only	Decision support systems	Statistical analyses in general
Abad-Grau et al. [4]	2008		✓	
Dreiseitl and Ohno-Machado [57]	2002	✓		
Harper [86]	2005	✓		
Kononenko [107]	2001	✓		
Lee and Abbott [117]	2003			✓
Ohno-Machado [149]	2001			✓
Podgorelec et al. [167]	2002	✓		
Scotch et al. [198]	2010			✓
Smith et al. [207]	2003		✓	

informatics journals instead of providing a review for a broader audience. The aim of this dissertation’s machine learning literature review is the discovery of relevant work in the field of, e.g., health care as well. Therefore, these nine reviews will be excluded from further examination. Moreover, only articles published since 2008 are categorized in order to give a picture of the recent developments in the interdisciplinary literature.

2.1.2.1 Attribute Ranking and Selection Techniques

Table 2.3 provides a categorization of the journal articles (from 2008–present) into different attribute ranking and selection technique categories. The table reveals that, the majority of the journal articles falls into the category “other attribute selection or evaluation techniques”. One explanation for this is that in some of these articles, new attribute selection techniques are developed which are compared with standard techniques. Correspondingly, one can observe that a number of articles are represented in both, the category “other attribute selection or evaluation techniques” and in categories such as “information gain”. Moreover, in some of these papers, the authors employ, e.g., regression models to discover relevant attributes. Alternatively or additionally, they decide based on expert opinions whether or not attributes are selected for the classification task. Besides this, many articles do not account for data preprocessing by selecting relevant attributes at all. Surprisingly, Markov blanket (MB) attribute selection was only discovered three times in the search for relevant publications although, it is a very useful hybrid approach for attribute selection and classification and is used in this dissertation to study attribute selection in connection with early DRG classification of inpatients.

Table 2.3 Overview of attribute ranking and selection methods employed in health care from 2008–present

Classification without attribute ranking/selection	[6, 7, 14, 17, 28, 39, 48, 50, 58, 72, 90, 97, 100, 109, 124, 129, 133, 139, 150, 164, 178, 187, 189–192, 202, 208, 210, 211, 215, 217, 220, 230, 237, 240, 245]
Correlation-based	[5, 66, 71, 73, 88, 148]
Information gain	[5, 9, 16, 66, 68, 91, 94, 99, 104, 128, 148]
Markov blanket	[16, 88, 148]
Other attribute selection or evaluation techniques	[5, 9, 11, 12, 15, 19, 25, 37, 38, 40, 41, 47, 49, 52, 56, 59–61, 66–68, 71, 73–75, 85, 88, 89, 94, 96, 98, 108, 110, 118, 122, 123, 128, 130, 132, 141–143, 146, 155–158, 168–170, 179, 180, 182, 183, 188, 209, 212, 221, 223, 225, 226, 232, 234–236, 241, 244, 246]
Principal component	[11, 119–121, 143, 213, 219]
Relief algorithms	[40, 41, 66, 148]
Wrapper	[40, 41, 43, 66, 71, 88, 91, 126, 127, 148, 153, 159, 161, 183, 214]

2.1.2.2 Classification Techniques

Table 2.4 provides a categorization of the journal articles into classification techniques. The table reveals that in most of the publications, support vector machines were employed, followed by other classification techniques and decision trees. On the contrary, decision rules as well as Bayesian networks, including Markov blanket represent the minority of the classification methods. Similar to the attribute selection techniques, the category “other classification techniques” contains, among others, innovative approaches that have been developed for specific classification tasks. One example for this is the work of Mu et al. [143] who employ principal component analysis as an attribute selection but employ discriminant analysis as a classification technique. Since the aim of this dissertation is to provide a general overview of standard methods, their application in health care and their evaluation for early DRG classification, the methods provided in the category “other classification techniques” are not examined in detail.

Among the articles, a number of publications deserve particular attention: Bai et al. [16], Fan and Chaovalitwongse [60] and Miettinen and Juhola [139]. These articles employ Bayesian models in order to pre-process data or to perform classification of medical diagnoses. Goodson and Jang [75] evaluate Bayesian networks in order to assess quality of care in the context of nursing home care.

Besides the articles classified above, further relevant publications are, e.g., Hall and Holmes [80] who study the connection between attribute selection and classification. They evaluate a sampling technique described by Robnik-Šikonja and Kononenko [184] and compare it with further attribute selection techniques using standard data sets. The attribute selection techniques are benchmarked

Table 2.4 Overview of classification methods employed in health care from 2008–present

Artificial neural networks	[11, 16, 19, 37, 43, 49, 50, 58, 73, 85, 88–90, 97, 99, 110, 118, 119, 121, 122, 124, 127, 129, 132, 133, 155, 169, 171, 209, 215, 219, 220, 223, 244]
Bayesian networks	[6, 16, 66, 75, 88, 122, 126, 133, 139, 150, 153, 210, 211, 246]
Combined classification	[6, 15, 28, 37, 39, 43, 52, 66, 74, 88, 89, 94, 97, 100, 109, 148, 153, 158, 171, 213, 220, 223, 226]
Decision rules	[6, 25, 52, 56, 61, 66, 91, 97, 99, 108, 123, 146, 148, 153, 158, 230, 232, 235]
Decision trees	[5, 6, 15, 17, 28, 37, 39, 43, 47–50, 66–68, 72, 74, 89–91, 94, 97, 99, 100, 104, 108, 109, 121, 122, 127, 129, 133, 153, 155, 157, 161, 168, 170, 171, 179, 180, 187, 189–192, 211, 214, 217, 226, 244, 246]
Markov blanket	[16, 88]
Naive Bayes	[6, 7, 12, 14, 16, 47, 48, 52, 59, 66, 68, 73, 74, 89, 97–99, 104, 109, 122, 124, 126, 132, 133, 139, 141, 148, 150, 153, 161, 182, 202, 211, 225, 230, 232]
Other classification techniques	[5, 6, 9, 16, 25, 38, 39, 43, 48, 60, 66–68, 71, 74, 75, 90, 91, 96–100, 104, 109, 110, 118, 121–124, 127, 129, 132, 133, 139, 141, 143, 153, 155, 158, 159, 161, 164, 168, 171, 178, 182, 183, 187–189, 212, 223, 225, 226, 234, 236]
Support vector machines	[5, 9, 16, 25, 39–41, 43, 47, 48, 50, 56, 60, 66, 68, 73, 88–90, 94, 97, 100, 109, 110, 120–122, 124, 128, 130, 132, 133, 141, 142, 148, 156, 158, 159, 164, 169, 178, 180, 182, 183, 187, 188, 190, 208, 212, 213, 221, 223, 234, 237, 240, 241, 244–246]

using two different classification methods. Methods that combine classification models are described by Kuncheva [111]. Similarly, in this dissertation, different attribute selection techniques that combine classifiers are evaluated. Ramiarina et al. [176] consider the prediction of a continuous attribute (costs) and thus employ a regression model. Similar to Bertsimas et al. [21] whose work was mentioned in the beginning (see Sect. 2.1.2), the study is excluded from the overview because the authors employ a regression model to predict health care costs. Instead, in this dissertation, a discrete attribute (DRG) is predicted and, therefore, we have a classification problem. Grubinger et al. [77] suggest the use of classification and regression trees in order to group inpatients with similar lengths of stay, instead of classifying individual patients. The results of their study are a selection of classification tree models and recommendations for the further development of the Austrian DRG system. Busse et al. [30] consider the problem of coding clinical data to the correct DRG from a quality management perspective. Coding quality highly influences classification and the assignment of the inpatient to the correct

DRG for billing purposes. However, Busse et al. [30] deal with the computation of DRGs after the discharge of the inpatient, whereas the problem addressed in this dissertation is to assign the inpatient to the appropriate DRG before and at admission for operations-driven DRG classification. In the next section, the methods employed for attribute selection and DRG classification are summarized.

2.2 Attribute Ranking and Selection Techniques Employed for Early DRG Classification

In what follows, a formal description of the early DRG classification problem is provided before the different approaches are presented. Let \mathcal{I} denote a set of individuals (hospital inpatients) and let \mathcal{D} denote the set of DRGs to which these individuals will be classified. For each inpatient $i \in \mathcal{I}$, we observe a set of attributes \mathcal{A} at the time the patient contacts the hospital for admission, while the inpatient's true DRG, $d_i \in \mathcal{D}$, is computed once the inpatient is discharged. Let \mathcal{V}_a denote the set of possible values for attribute $a \in \mathcal{A}$ and let $v_{i,a} \in \mathcal{V}_a$ denote the value of attribute a for inpatient i . As soon as inpatient i is admitted to the hospital, given the inpatient's values $v_{i,a}$ for each attribute $a \in \mathcal{A}$ the objective is to predict d_i . In this supervised learning problem, it is assumed that labeled training data from many other inpatients $j \in \mathcal{I} \setminus i$ is available in which attribute values $v_{j,a}$ and DRGs d_j are known. This training data is used to learn a classification model, which is then used for DRG prediction.

As indicated by the name, attribute ranking techniques provide an ordered list of attributes while attribute selection techniques select from all available attributes a subset of attributes which are relevant for classification. In this study, the following attribute selection techniques are considered: Information gain (IG), Relief-F attribute ranking (see Hall and Holmes [80]), Markov blanket attribute selection (see Bai et al. [16]), correlation-based feature selection (CFS) as well as wrapper attribute selection.

2.2.1 Information Gain Attribute Ranking

In order to describe the IG attribute ranking technique, the concept of information entropy is introduced. The idea is well-known from information theory and it measures the uncertainty associated with an attribute (see Sharma and Yu [205]). Given the prior probability $p(d)$ for each DRG $d \in \mathcal{D}$, the information entropy $H(\mathcal{D})$ is defined by Eq. (2.1).

$$H(\mathcal{D}) = - \sum_{d \in \mathcal{D}} p(d) \ln p(d) \quad (2.1)$$

Table 2.5 An example set of instances

i	d_i	$v_{i,1}$	$v_{i,2}$	$v_{i,3}$
1	I74C	Male	0–30	Fracture
2	I74C	Male	0–30	Fracture
3	I74C	Female	0–30	Fracture
4	I74C	Male	0–30	Chest pain
5	F62A	Male	0–30	Chest pain
6	F62A	Female	0–30	Pneumonia
7	F62A	Female	31–100	Pneumonia
8	F62A	Female	31–100	Pneumonia
9	F62A	Male	31–100	Pneumonia
10	F62A	Male	31–100	Heart failure
11	F62C	Female	31–100	Heart failure
12	F62C	Male	31–100	Heart failure

It holds for the discrete case and in the case that $p(d) = 0$, the convention is that $0 \ln(0) \equiv 0$ since $\lim_{x \rightarrow 0^+} x \ln x = 0$ (see Mackay [131]). The negative sign ensures that $H(\mathcal{D})$ is positive or zero and the more uniform an attribute value is distributed over all instances, the higher is its entropy (see Bishop [22]). Using Eq. (2.2) one can compute the conditional information entropy $H(\mathcal{D}|a)$ of \mathcal{D} , given an attribute $a \in \mathcal{A}$. Here, $p(v)$ is the prior probability of attribute value $v \in \mathcal{V}_a$ for attribute $a \in \mathcal{A}$ and $p(d|v)$ is the conditional prior probability of a DRG d , given an attribute value $v \in \mathcal{V}_a$ of attribute $a \in \mathcal{A}$.

$$H(\mathcal{D}|a) = - \sum_{v \in \mathcal{V}_a} p(v) \sum_{d \in \mathcal{D}} p(d|v) \ln p(d|v) \quad (2.2)$$

The information gain $IG(a)$ of each attribute $a \in \mathcal{A}$ is then computed by employing Eq. (2.3).

$$IG(a) = H(\mathcal{D}) - H(\mathcal{D}|a) \quad (2.3)$$

The higher the information gain $IG(a)$ of an attribute $a \in \mathcal{A}$, the more valuable the attribute is assumed to be for classifying \mathcal{D} . In the following, the IG concept is illustrated using an example with the set of attributes $\mathcal{A} := \{\text{gender, age, primary diagnosis}\}$, the set of DRGs $\mathcal{D} := \{\text{I74C, F62A, F62C}\}$ as well as the sets of attribute values $\mathcal{V}_{\text{gender}} := \{\text{male, female}\}$, $\mathcal{V}_{\text{age}} := \{0\text{--}30, 31\text{--}100\}$ and $\mathcal{V}_{\text{primary diagnosis}} := \{\text{fracture, chest pain, pneumonia, heart failure}\}$. Twelve sample instances are provided by Table 2.5.

In this table, column $v_{i,1}$ contains the values of attribute “gender”, $v_{i,2}$ the values of attribute “age” and $v_{i,3}$ the values of attribute “primary diagnosis”. Summary statistics are provided by Table 2.6.

Table 2.6 Summary statistics of gender, age, primary diagnosis and DRG from the example provided by Table 2.5

	I74C	F62A	F62C
Gender			
Male	3	3	1
Female	1	3	1
Age			
0–30	4	2	0
31–100	0	4	2
Primary diagnosis			
Fracture	3	0	0
Chest pain	1	1	0
Pneumonia	0	4	0
Heart failure	0	1	2
DRG			
	4	6	2

Applying Eq.(2.1) to Table 2.6, the information entropy $H(\mathcal{D})$ of the class DRG \mathcal{D} comes up to $H(\mathcal{D}) = -(\frac{4}{12} \ln \frac{4}{12} + \frac{6}{12} \ln \frac{6}{12} + \frac{2}{12} \ln \frac{2}{12}) = 1.011$. Using Eq.(2.2) and the conditional frequencies shown in Table 2.6, the conditional information entropy for the attribute “gender” can be computed as follows: $H(\mathcal{D}|\text{gender}) = -(\frac{7}{12}(\frac{3}{12} \ln \frac{3}{12} + \frac{3}{12} \ln \frac{3}{12} + \frac{1}{12} \ln \frac{1}{12}) + \frac{5}{12}(\frac{1}{12} \ln \frac{1}{12} + \frac{3}{12} \ln \frac{3}{12} + \frac{1}{12} \ln \frac{1}{12})) = 0.842$. Accordingly, the conditional entropies of the attributes “age” and “primary diagnosis” come up to $H(\mathcal{D}|\text{age}) = 0.664$ and $H(\mathcal{D}|\text{primary diagnosis}) = 0.404$. Using Eq.(2.3), the IG for the attribute “gender” is $IG(\text{gender}) = H(\mathcal{D}) - H(\mathcal{D}|\text{gender}) = 1.011 - 0.842 = 0.169$ and for “age” and “primary diagnosis”, the information gains come up to $IG(\text{age}) = 0.347$ and $IG(\text{primary diagnosis}) = 0.607$, respectively. The results reveal that the attribute “primary diagnosis” has the highest IG, “age” has the second highest and “gender” has the lowest IG.

Since IG considers each attribute individually, and thus is ill-suited for examining the potential contribution of attribute combinations, the use of a further attribute ranking technique (Relief-F) will be examined.

2.2.2 Relief-F Attribute Ranking

Relief algorithms are known as fast feature selection algorithms (see Aliferis et al. [8]). Kira and Rendell [102] have developed this class of algorithms which has shown to be very efficient for binary classification problems (see Robnik-Šikonja and Kononenko [184]). The original algorithm has been refined by Robnik-Šikonja and Kononenko [184]: Their Relief-F variant is evaluated in this dissertation because, in contrast to Relief, it is not limited to two class problems and can deal with incomplete and noisy data (see Robnik-Šikonja and Kononenko [184]).

In order to describe the algorithm, the “ k -nearest hits” and “ k -nearest misses” for a sampled instance $i \in \mathcal{I}$ have to be defined. Let the set of k -nearest hits $\mathcal{H}_i(k) \subset \mathcal{I} \setminus i$ of an instance $i \in \mathcal{I}$ contain at most k instances $j \in \mathcal{I}, j \neq i$ which have the same DRG d_i as instance i . Specifically, those instances j are chosen with the same DRG such that $d_j = d_i$ and which have the lowest $\text{diff}_{i,j}$ -values as defined by Eqs. (2.4) and (2.5).

$$\text{diff}_{i,j} = \sum_{a \in \mathcal{A}} \text{diff}_{i,j,a} \quad (2.4)$$

$$\text{diff}_{i,j,a} = \begin{cases} 0, & \text{if } v_{i,a} = v_{j,a} \\ 1, & \text{otherwise} \end{cases} \quad (2.5)$$

Furthermore, for each DRG $d \neq d_i$, let the set of k -nearest misses $\mathcal{M}_{d,i}(k) \subset \mathcal{I} \setminus i$ of instance i contain at most k instances $j \in \mathcal{I}, j \neq i$ with $d_j = d$ which have the lowest $\text{diff}_{i,j}$ -values as defined by Eqs. (2.4) and (2.5). Both the k -nearest hits and the k -nearest misses for each DRG $d \in \mathcal{D}$ are inserted into Eq. (2.6) which computes the quality measure Q_a for attribute $a \in \mathcal{A}$.

$$Q_a = \frac{1}{k \cdot |\mathcal{I}|} \sum_{i \in \mathcal{I}} \left(- \sum_{h \in \mathcal{H}_i(k)} \text{diff}_{i,h,a} + \sum_{d \in \mathcal{D} \setminus d_i} \frac{p(d)}{1 - p(d_i)} \sum_{m \in \mathcal{M}_{d,i}(k)} \text{diff}_{i,m,a} \right) \quad (2.6)$$

For large data sets, this computation can be time-consuming. Therefore, an adaptation of this equation is provided which is employed in the Relief-F sampling algorithm while s denotes the number of samples. For attribute $a \in \mathcal{A}$ and a sampled instance $i \in \mathcal{I}$ the quality $Q_{a,i}$ is computed by Eq. (2.7), see Robnik-Šikonja and Kononenko [184].

$$Q_{a,i} = \frac{1}{k \cdot s} \left(- \sum_{h \in \mathcal{H}_i(k)} \text{diff}_{i,h,a} + \sum_{d \in \mathcal{D} \setminus d_i} \frac{p(d)}{1 - p(d_i)} \sum_{m \in \mathcal{M}_{d,i}(k)} \text{diff}_{i,m,a} \right) \quad (2.7)$$

Equation (2.7) is used in the sampling method, described by Algorithm 1. Here, $w(a)$ is a quality measure for each attribute $a \in \mathcal{A}$ updated in each iteration.

In what follows, the algorithm is illustrated with the example from Table 2.5 in which the user-defined parameters are set to $s = 5$ random samples and $k = 2$ “nearest-neighbors”. Initially, the weights $w(a)$ are set for all attributes $a \in \mathcal{A}$ to 0 (see line 1). Afterwards, instance i is randomly selected from the data set (see line 3), e.g. instance $i = 3$ with DRG $d_3 = \text{I74C}$. Computing the $\text{diff}_{3,j}$ -values for all instances $j \in \mathcal{I} \setminus 3$ that have the same DRG as instance $i = 3$ leads to the following results: $\text{diff}_{3,1} = 1, \text{diff}_{3,2} = 1, \text{diff}_{3,4} = 2$. Thus, the set of the 2-nearest hits (see line 4) is $\mathcal{H}_3(2) = \{1, 2\}$. Accordingly, the sets of the 2-nearest misses (see line 6) $\mathcal{M}_{d,3}(2)$ are $\{5, 6\}$ and $\{11, 12\}$ for $d = \text{F62A}$ and $d = \text{F62C}$, respectively. Now, for each attribute $a \in \mathcal{A}$ the attribute weight $w(a)$ is updated (see line 9).

Algorithm 1 Relief-F algorithmInput parameters: k, s Output parameters: $w(a) \quad \forall a \in \mathcal{A}$

```

1:  $w(a) := 0 \quad \forall a \in \mathcal{A}$ 
2: for  $l = 1$  to  $s$  do
3:   Randomly select an instance  $i \in \mathcal{I}$ 
4:    $\mathcal{H}_i(k) := k$ -nearest hits in  $\mathcal{I}$  based on instance  $i$ 
5:   for all  $d \in \mathcal{D} \setminus d_i$  do
6:      $\mathcal{M}_{d,i}(k) := k$ -nearest misses in  $\mathcal{I}$  based on DRG  $d$  and instance  $i$ 
7:   end for
8:   for all  $a \in \mathcal{A}$  do
9:      $w(a) := w(a) + Q_{a,i}$ 
10:  end for
11: end for

```

Table 2.7 Sample-dependent sets and parameters as well as each attribute's current weight for each iteration of Relief-F

l	Sample-dependent sets and parameters								Attributes current weight		
	i	d_i	H_i	$M_{1,i}(2)$	$M_{2,i}(2)$	$Q_{1,i}$	$Q_{2,i}$	$Q_{3,i}$	$w(1)$	$w(2)$	$w(3)$
1	3	174C	{1,2}	{5,6}	{11,12}	-0.100	0.050	0.200	-0.100	0.050	0.200
2	8	F62A	{7,9}	{3,4}	{11,12}	0.000	0.133	0.200	-0.100	0.183	0.400
3	12	F62C	{11}	{1,2}	{9,10}	-0.100	0.120	0.160	-0.200	0.303	0.560
4	7	F62A	{6,8}	{1,3}	{11,12}	0.100	0.033	0.200	-0.100	0.336	0.760
5	3	174C	{1,2}	{5,6}	{11,12}	-0.100	0.050	0.200	-0.200	0.386	0.960

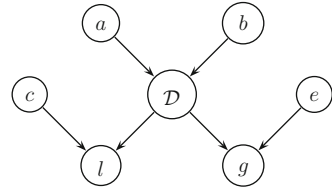
We start in iteration $l = 1$ and attribute $a = 1$ (gender). We compute the prior probabilities for the DRGs which are $p(d = 174C) = 0.333$, $p(d = F62A) = 0.500$ and $p(d = F62C) = 0.167$. Inserting these values into Eq.(2.6)

results in $Q_{1,3} = \frac{1}{2 \cdot 5} \left(- \sum_{h \in \mathcal{H}_3(2)} \text{diff}_{3,h,1} + \sum_{d \in \mathcal{D} \setminus d_3} \frac{p(d)}{1-p(d_3)} \sum_{m \in \mathcal{M}_{d,3}(2)} \text{diff}_{3,m,1} \right) = \frac{1}{10} \cdot \left(-(1+1) + \frac{\frac{6}{12}}{1-\frac{4}{12}} \cdot (1+0) + \frac{\frac{2}{12}}{1-\frac{4}{12}} \cdot (0+1) \right) = -0.100$. Accordingly, $w(1) = -0.100$. The results of the example are provided in Table 2.7.

Table 2.7 reveals that instance $i = 3$ can be sampled twice from the set of instances. Robnik-Šikonja and Kononenko [184] do not specify whether sampling an instance more than once is allowed. The last row in Table 2.7 reveals that attribute 3 has the highest attribute weight and attribute 1 the lowest. This leads to the suggestion that attribute 3 (primary diagnosis) is the most relevant for classification.

So far two methods that basically compute a weight for each attribute with respect to the class \mathcal{D} have been presented. The $IG(a)$ or $w(a)$ values of each attribute $a \in \mathcal{A}$ can be sorted by decreasing order (ranked). Then, the attributes with highest values can be selected for classification. As stated by Yu and Liu [239], a major drawback of the two methods presented so far is that they are not capable of detecting redundant attributes. This, however, can be overcome by using Markov blanket attribute selection, which will be introduced next.

Fig. 2.2 Markov blanket of vertex \mathcal{D}



2.2.3 Markov Blanket Attribute Selection

In order to introduce Markov blanket attribute selection, the notation of Bayesian networks (BN), a type of probabilistic graphical model will be introduced. A Bayesian network is a directed acyclic graph (DAG) $\mathcal{G} := (\mathcal{V}, \mathcal{E})$ with vertices \mathcal{V} and edges \mathcal{E} . In this graph, the vertices represent variables and the edges encode the conditional independence relationships between these variables (each variable is conditionally independent of its non-descendants in the graph given its parents). Pearl [160] and Wasserman [231] provide further theoretical properties of Bayesian networks and other probabilistic graphical models. For an overview of statistical graphical models with applications in systems biology, see Nagarajan et al. [144] as well as Scutari and Strimmer [200]. The Markov blanket of a vertex $v \in \mathcal{V}$, denoted by $MB(v)$, is a minimal subset of vertices containing vertex v , its direct parents and direct children as well as all direct parents of the children of v . The Markov blanket of vertex v contains all the variables needed to predict the value of that variable, since v is conditionally independent of all other variables given its Markov blanket. Figure 2.2 shows a sample Markov blanket DAG.

All vertices in the graph are part of the Markov blanket of vertex \mathcal{D} , since a and b are direct parents of \mathcal{D} , l and g are direct children of \mathcal{D} and c and e are direct parents of the children of \mathcal{D} .

In this application of DRG classification, the vertices of the graph include the DRG variable (\mathcal{D}) as well as all attributes $a \in \mathcal{A}$. The aim is to select the subset of attributes which are relevant for predicting \mathcal{D} and thus to be able to select those and only those variables in the Markov blanket of \mathcal{D} . Many methods have been developed to infer the Markov blanket of a variable from data, as described in Bai et al. [16] and Fu and Desmarais [69]. In order to describe the general procedure to derive a Markov blanket from data, the algorithm devised by Ramsey et al. [177] will be described because it can be illustrated straightforwardly with a simple example. With its breadth-first search strategy, the algorithm initially tries to connect all variables (many are similar, e.g. diagnoses) with the target variable DRG conditioned on a limited and therefore small subset of variables. The algorithm can be adapted for the discovery of the Markov blanket of a DRG as follows: The input parameters are the set of attributes \mathcal{A} , the set of DRGs \mathcal{D} and a maximum search depth d_{\max} . The output is the Markov blanket of the target variable DRG, represented by the graph \mathcal{G} .

Algorithm 2 Markov blanket search heuristicInput parameters: Set of attributes \mathcal{A} , set of DRGs \mathcal{D} , maximum search depth d_{\max} Output parameter: Markov blanket $\mathcal{G} := (\mathcal{V}, \mathcal{E})$

```

1:  $\mathcal{V} := \emptyset, \mathcal{E} := \emptyset, sepSet := \emptyset, Forbidden := \emptyset$ 
2:  $\mathcal{V} := \mathcal{D} \cup \mathcal{A}$ 
3:  $adj(\mathcal{D}) := checkedges(\mathcal{D})$ 
4: for all  $a \in adj(\mathcal{D})$  do
5:    $adj(a) := checkedges(a)$ 
6: end for
7: for all  $b \in adj(a) \setminus \{adj(\mathcal{D}) \cup \mathcal{D}\}$  do
8:    $adj(b) := checkedges(b)$ 
9: end for
10:  $\mathcal{G} := orient-edges(\mathcal{G})$ 
11:  $\mathcal{G} := trim-to-Markov-blanket(\mathcal{G})$ 

```

Algorithm 3 Checkedges (see Ramsey [177])Input parameters: Vertex a , graph $\mathcal{G} := (\mathcal{V}, \mathcal{E})$, depth of search d_{\max} , set mapping $sepSet(a, b)$, set of edges *Forbidden*Output parameters: Updated graph $\mathcal{G} := (\mathcal{V}, \mathcal{E})$, updated set mapping $sepSet(a, b)$, updated set of edges *Forbidden*

```

1: for all  $b \in \mathcal{V} \setminus a$  do
2:   if  $\{a, b\} \notin Forbidden$  then  $\mathcal{E} := \mathcal{E} \cup \{a, b\}$ 
3:   end if
4: end for
5: for  $depth = 0, \dots, d_{\max}$  do
6:   if  $|adj(a)| \geq depth + 1$  then
7:     for all  $b \in adj(a)$  do
8:       if  $b \perp\!\!\!\perp a | \mathcal{S}$  given  $\mathcal{S} \subset \{adj(a) \setminus b\} : |\mathcal{S}| = depth$  then
9:          $\mathcal{E} := \mathcal{E} \setminus \{a, b\}$ 
10:         $Forbidden \cup \{a, b\}$ 
11:         $sepSet\{a, b\} = \mathcal{S}$ 
12:      end if
13:    end for
14:   end if
15: end for

```

Here, $adj(a)$ denotes the set of adjacent vertices of the vertex a . For testing conditional independence, the χ^2 -test can be used with an $\alpha = 5\%$ level of significance. The result of this test ($a \perp\!\!\!\perp b | \mathcal{A}$) denotes that attribute $a \in \mathcal{A}$ is conditionally independent of attribute $b \in \mathcal{A} \setminus a$ given further attributes $\mathcal{A} : a, b \notin \mathcal{A}$. The set of unordered pairs $\{a, b\} \in Forbidden$ denotes forbidden edges in the graph, determined by the function *checkedges* which is described by Algorithm 3.

An edge is forbidden if a subset of attributes $\mathcal{S} \subset \mathcal{V} \setminus \{a, b\}$ exists that separates a from b . For each edge $\{a, b\} \in Forbidden$ a set mapping $sepSet(a, b) = \mathcal{S}$ exists which in turn contains \mathcal{S} . $sepSet(a, b)$ and *Forbidden* are updated synchronously

Algorithm 4 Orient-edges (see Ramsey [177]). For the definition of a collider, see e.g. Wasserman [231]. The asterisk (*) can be the head or the tail of an arc such that e.g. $a * -b$ can denote (b, a) or $\{a, b\}$

Input parameter: $\mathcal{G} := (\mathcal{V}, \mathcal{E})$

Output parameter: $\mathcal{G} := (\mathcal{V}, \mathcal{E})$

```

1: for all triples of vertices  $\{a, b, c\} \in \mathcal{V}$  do
2:   if  $a * -b - *c, a \notin \text{adj}(c), b \notin \text{sepSet}(a, c)$  then orient  $a * \rightarrow b \leftarrow *c$ .
3:   end if
4: end for
5: for all 4-tuples of vertices  $\{a, b, c, d\} \in \mathcal{V}$  do
6:   if  $a \rightarrow b, b - c, a \notin \text{adj}(c), \neg(a \rightarrow b \leftarrow c)$  then  $b \rightarrow c$ .
7:   end if
8:   if  $a \rightarrow b, b \rightarrow c, a - c$  then  $a \rightarrow c$ .
9:   end if
10:  if  $a - b, a - c, a - d, c \rightarrow b, d \rightarrow b$  then  $a \rightarrow b$ .
11:  end if
12:  if  $a - b, b \in \text{adj}(d), a \in \text{adj}(c), a - d, b \rightarrow c, c \rightarrow d$  then  $a \rightarrow d$ .
13:  end if
14: end for

```

by *checkedges* where $\text{adj}(a)$ denotes the set of adjacent vertices of the vertex a . Orient-edges orients edges to arcs and is presented in Algorithm 4.

After the initialization in line 1, the algorithm sets the set of vertices equal to DRG and the set of attributes. *checkedges* adds for all attributes $a \in \mathcal{A}$ an edge from a to \mathcal{D} such that $\mathcal{E} := \mathcal{E} \cup \{a, \mathcal{D}\}$. Next, the algorithm tests, conditioned on subsets of the vertices in the graph, whether or not an edge $\{a, \mathcal{D}\}$ can be removed from \mathcal{E} . An edge $\{a, \mathcal{D}\}$ is removed if there is a subset $\mathcal{S} \subset \mathcal{V} \setminus \{a, \mathcal{D}\}$ of vertices in the graph such that $(a \perp\!\!\!\perp \mathcal{D} | \mathcal{S})$. Here, $|\mathcal{S}|$ is limited by the user-defined parameter d_{\max} , which is the maximum search depth. Depending on d_{\max} and the speed of the conditional independence test, this procedure can be time-consuming. Next, the same procedure is performed with adjacents of \mathcal{D} and adjacents of the adjacents of \mathcal{D} . Note that until line 9 in Algorithm 2, we search an undirected graph. After that, Algorithm 4 orients the edges to arcs and different orientation rules are applied while ensuring that the property of acyclicity is retained (see Meek [137] and Bai et al. [16]) until no further orientation can be made. Finally, when reaching line 11 of Algorithm 2 the graph \mathcal{G} is trimmed to the Markov blanket of \mathcal{D} so that a Markov blanket DAG is returned.

In what follows, an example is provided by employing the publicly available data set “learning.test” containing discrete attributes (see Scutari [199]). It consists of the set of attributes $\mathcal{A} = \{A, \dots, F\}$. The causal network including the Markov blanket of attribute A is provided in Fig. 2.3.

Assume, the Markov blanket of the attribute A has to be inferred. Let the results of the conditional independence tests during the execution of *checkedges* be: $(C \perp\!\!\!\perp A | \emptyset), (F \perp\!\!\!\perp A | \emptyset), (E \perp\!\!\!\perp A | B), (C \perp\!\!\!\perp B | \emptyset), (F \perp\!\!\!\perp B | \emptyset), (D \perp\!\!\!\perp B | A), (F \perp\!\!\!\perp D | \emptyset), (E \perp\!\!\!\perp D | A), (C \perp\!\!\!\perp E | \emptyset), (F \perp\!\!\!\perp C | \emptyset)$. Figure 2.4 shows the different steps of the algorithm

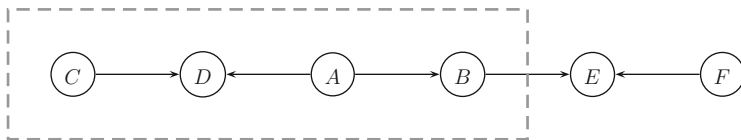


Fig. 2.3 Causal network of the example (see Scutari [199]) including the Markov blanket of attribute A (dashed gray rectangle)

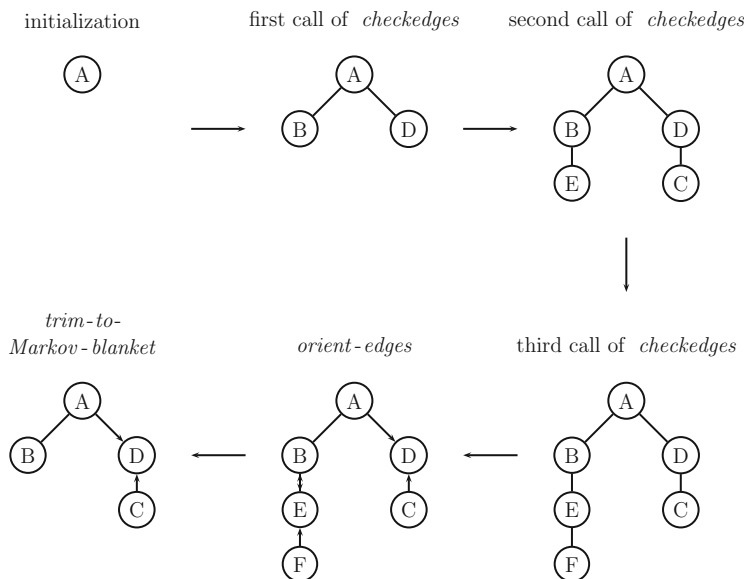


Fig. 2.4 Different steps during the Markov blanket search

when using $d_{\max} = 1$, the χ^2 -test as used by Scutari [199] and a confidence level of $\alpha = 5\%$. Table 2.8 shows the sets after the termination of the algorithm.

Note that in the *orient-edges*-step, the edge $\{A, B\}$ cannot be oriented since B is an element of the set that separates A from E (see line 2 of Algorithm 4). Because of the Markov condition in undirected graphs (see Wasserman [231]), B separates E and F from A such that the Markov blanket of A is $MB(A) = \{A, B, C, D\}$ which is correct (cf. Fig. 2.3).

Besides Ramsey's [177] algorithm, many other methods have been developed to obtain the Markov blanket of a variable from data. In this dissertation, two of them will be evaluated: The so-called Grow-Shrink approach (GS) devised by Margaritis [135] and the Incremental-Association search (IA) devised by Tsamardinos et al. [218]. The reason to employ these algorithms is because arcs can be fixed in the Markov blanket DAG of which a functional relationship between attributes and DRG are known as required by the DRG system (see Fig. 1.1). Hence, it can be controlled that these attributes are not considered irrelevant or redundant.

Table 2.8 Sets after the third call of the *checkedges* algorithm to find adjacents of the adjacents of the vertex A

Set	Elements
\mathcal{V}	A, B, C, D, E, F
\mathcal{E}	$\{A, B\}, \{A, D\}, \{E, B\}, \{C, D\}, \{F, E\}$
<i>Forbidden</i>	$\{A, C\}, \{A, F\}, \{A, E\}, \{B, D\}, \{B, C\}, \{B, F\}, \{F, D\}, \{E, D\}, \{C, E\}, \{C, F\}$
$sepSet(A, C)$	\emptyset
$sepSet(A, F)$	\emptyset
$sepSet(A, E)$	B
$sepSet(B, D)$	A
$sepSet(B, C)$	\emptyset
$sepSet(B, F)$	\emptyset
$sepSet(F, D)$	\emptyset
$sepSet(E, D)$	A
$sepSet(C, E)$	\emptyset
$sepSet(C, F)$	\emptyset

2.2.4 Correlation-Based Feature Selection

Another way to select attributes is to select ones that individually correlate well with the class (DRG) and have low intercorrelation with other individual attributes. In order to compute the intercorrelation of two nominal attributes a and b , nominal attributes are the majority of attributes evaluated in this dissertation (see Table B.1), one has to compute the symmetrical uncertainty $U(a, b) \in [0; 1]$ by employing the following equation (see, e.g., Hall and Holmes [80]):

$$U(a, b) = 2 \cdot \frac{H(a) + H(b) - H(a|b)}{H(a) + H(b)}. \quad (2.8)$$

Again, $H(a)$ is the entropy of attribute a (see Eq. (2.1)) while $H(a|b)$ is the conditional entropy of attribute a given attribute b using Eq. (2.2). The attribute subset \mathcal{A}_i^* which maximizes the following expression is selected:

$$\mathcal{A}_i^* = \arg \max_{\mathcal{A}' \subset \mathcal{A}} \frac{\sum_{a \in \mathcal{A}'} U(a, \mathcal{D})}{\sqrt{\sum_{a \in \mathcal{A}'} \sum_{b \in \mathcal{A}' \setminus a} U(a, b)}}. \quad (2.9)$$

2.2.5 Wrapper Attribute Selection

In what follows, a method will be described that “wraps” a classification scheme into the attribute selection procedure. For this attribute subset evaluation, a classification scheme as well as an evaluation measure have to be chosen that will be optimized, e.g. accuracy (Acc.). The approach is described in Table 2.9 using $\mathcal{A} := \{a, b, c\}$ as a set of attributes (for details, see Kohavi and John [105]).

Table 2.9 Wrapper attribute subset evaluation in order to produce a ranked list of attributes

Iteration 1			Iteration 2			Iteration 3		
Attribute set	Acc.	Best attribute	Attribute set	Acc.	Best attribute	Attribute set	Acc.	Best attribute
<i>a</i>	0.1		<i>a b</i>	0.3		<i>a b c</i>	0.35	
<i>b</i>	0.3	<i>b</i>	<i>b c</i>	0.4	<i>c</i>	<i>b c</i>	0.4	–
<i>c</i>	0.2							

Starting with an empty subset of attributes, in each iteration one (best) single attribute is added to the list of attributes. In the example, we choose in the first iteration attribute *b* since it has the highest gain in accuracy (see Table 2.9). In the second iteration (see Table 2.9) we check whether attribute *a* or *c* can improve classification accuracy. Since the additional attribute *c* results in the highest increase of accuracy, it is added to the set of attributes. Finally, based on attributes *b* and *c* during the third iteration (see Table 2.9) accuracy is evaluated again to check whether attribute *a* can improve accuracy. Since accuracy cannot be improved, the subset $\{b, c\} \subset \mathcal{A}$ is selected as the best subset of attributes. Usually, this greedy search goes along with high computational effort which depends on the complexity of the classification scheme and on the number of attributes, among others.

2.3 Classification Techniques Employed for Early DRG Classification

In the following, six classification methods will be summarized: Naive Bayes (NB), Bayesian networks (BN), classification trees (also called decision trees), a method that combines these classifiers by voting, a probability averaging approach (PA) and a straightforward decision rule. For each method, the classifier is learned from a dataset of labeled training examples. This means that the true DRG of each inpatient is known to the classification method. Afterwards, the classifier is applied to a separate dataset of unlabeled test examples. Here, the true DRG of each inpatient is unknown to the classification method and must be predicted. The naive Bayes and Bayesian network methods infer a probabilistic model from the training data and compute the posterior probability that the inpatient belongs to a DRG *d* given the inpatient’s attributes \mathcal{A} . Then, the inpatient is assigned to that DRG with the highest posterior probability. The classification tree method, instead, infers a tree-structured set of decision rules from the training data and uses these rules to predict the inpatient’s DRG. The voting and the probability averaging approaches assign the patient to that DRG which receives the highest support based on the individual classification methods. Finally, a decision-rule based approach is presented, based on historical data. In the following, each method will be described in more detail.

2.3.1 Naive Bayes

The naive Bayes classifier assumes that all of an inpatient's attributes $a \in \mathcal{A}$ are conditionally independent, given the inpatient's DRG d . Under this assumption, the classifier assigns the DRG d_i^* to the test instance i employing Eq. (2.10).

$$d_i^* = \arg \max_{d \in \mathcal{D}} \left\{ p(d) \prod_{a=1}^{|\mathcal{A}|} p(v_{i,a}|d) \right\} \quad (2.10)$$

The prior probability $p(d)$ of each DRG d is learned from the training data by maximum likelihood estimation, i.e., $p(d)$ is set equal to the proportion of training examples which belong to class d . Similarly, the conditional likelihood of each attribute value $v_{i,a}$ given each DRG d is learned from the training data by maximum likelihood estimation, i.e., $p(v_{i,a}|d)$ is set equal to the proportion of training examples of class d which have value $v_{i,a}$ for attribute a .

In what follows, data from Table 2.5 is employed to provide an example for the naive Bayes classification. The data set is split into one test instance (instance 1) and 10 training instances (instances 2–11). The prior probabilities for each DRG in the training set are $p(I74C) = 0.273$, $p(F62A) = 0.545$ and $p(F62C) = 0.182$. Employing Eq. (2.10) the conditional probability for predicting the DRG of instance $i = 1$ and $d = I74C$ is $p(I74C) \cdot p(\text{male}|I74C) \cdot p(0-30|I74C) \cdot p(\text{fracture}|I74C) = 0.273 \cdot 0.182 \cdot 0.273 \cdot 0.182 = 0.002$. This is the maximum conditional probability for all DRGs $d \in \mathcal{D}$ so that the DRG of instance $i = 1$ is classified to $d_1^* = I74C$ which is correct.

This way of computing conditional probabilities is ill-suited to predicting the DRG, when events have not yet been observed. In the case of Tables 2.5 and 2.6, for example, the conditional probability of $p(F62C|0-30) = 0$. Accordingly, for Eq. (2.10), with any DRG d of which its conditional attribute value $v_{i,a}$ has not been observed, the entire conditional probability would become zero. To overcome this problem of underestimation, a Laplace estimator is commonly employed by simply adding 1 to each count (see Witten and Frank [233]).

2.3.2 Bayesian Networks

The naive Bayes approach assumes that each attribute is only dependent on the DRG but not dependent on other attributes, which is rarely true. Thus, the naive Bayes classifier is extended to a Bayesian network classifier, in which the set of conditional independence assumptions is encoded in a Bayesian network as described above. As in the naive Bayes approach, conditional probabilities are inferred from the training data, but now we must condition not only on the DRG d , but also on any other parents Π_a of the given attribute a in the Markov blanket graphical model:

$$d_i^* = \arg \max_{d \in \mathcal{D}} \left\{ p(d) \prod_{a=1}^{|\mathcal{A}|} p(v_{i,a} | d, \Pi_a) \right\} \quad (2.11)$$

Similar to the naive Bayes approach, the instance is assigned to that DRG d_i^* which has the highest posterior probability, as in the naive Bayes approach.

Given the graph in Fig. 2.2 an estimator of the conditional probabilities can be obtained by computing conditional probabilities of the vertices in the Markov blanket. Let us assume that the set of DRGs consists of d_1 and d_2 . Further, assume that an instance $i = 1$ with attribute values $v_{1,a}, v_{1,b}, v_{1,c}, v_{1,e}, v_{1,l}, v_{1,g}$ has to be classified. We also assume that we have test instances that have been used to compute the necessary conditional probabilities that can be derived from the graph. To classify the instance, we compute $p(\mathcal{D} = d_1 | v_{1,a}, v_{1,b}, v_{1,c}, v_{1,e}, v_{1,l}, v_{1,g}) = p(a = v_{1,a} | d_1) \cdot p(b = v_{1,b} | d_1) \cdot p(l = v_{1,l}, c = v_{1,c} | d_1) \cdot p(g = v_{1,g}, e = v_{1,e} | d_1)$ and $p(\mathcal{D} = d_2 | v_{1,a}, v_{1,b}, v_{1,c}, v_{1,e}, v_{1,l}, v_{1,g}) = p(a = v_{1,a} | d_2) \cdot p(b = v_{1,b} | d_2) \cdot p(l = v_{1,l}, c = v_{1,c} | d_2) \cdot p(g = v_{1,g}, e = v_{1,e} | d_2)$. Finally, the instance is assigned to that DRG which has the highest probability. As for naive Bayes, for attribute values in the test instance that have not yet occurred in the set of training instances, the conditional probability would be zero. In this case, a nearest-neighbor heuristic (see e.g. Bai et al. [16]) can be used.

2.3.3 Classification Trees

As stated in Sect. 1.1, the hospital where this study was undertaken employs a DRG grouper to determine the DRG of an inpatient from the second day after admission. The algorithm which is implemented in this software is similar to a classification tree: It is a tree-structured set of rules which deterministically computes each inpatient's DRG given their attribute values. In the context of this dissertation, a classification tree is a hierarchical data structure that consists of a root node which represents an attribute. Additional nodes that represent further attributes, except the "root attribute", are linked with the root node directly or indirectly. Leaf nodes represent the DRGs. Arcs between nodes represent the values of the attributes located in the predecessor hierarchy. A sample classification tree is shown in Fig. 2.5.

Instead of using a pre-existing set of decision rules as employed by the DRG grouper, we learn the classification tree automatically from the labeled training dataset. There are various methods to learn the structure of a classification tree from data: In this dissertation, Quinlan's algorithm [175] will be evaluated which has also been investigated by Hall and Holmes [80] with respect to attribute selection. The advantage of employing this approach is that the over-fitting of the classification tree as well as the tree size during the learning process can be controlled.

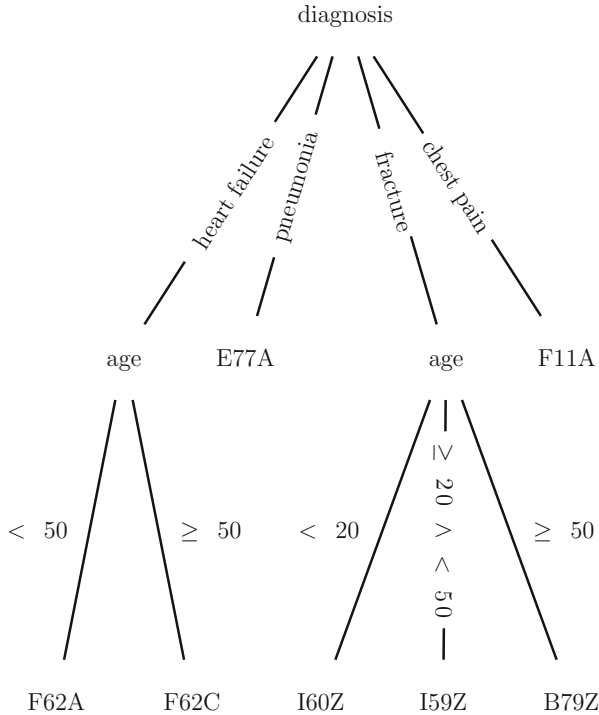


Fig. 2.5 Sample classification tree

The algorithm works as follows: In the first step, the attribute a^* with the maximum information gain is selected from the set of attributes \mathcal{A} . Based on a^* , which becomes the root node, \mathcal{I} is divided into subsets \mathcal{I}_v ; each one contains different values $v \in \mathcal{V}_{a^*}$ of attribute a^* . Each value is represented by an edge. If in any subset \mathcal{I}_v only one DRG d exists, the attribute value v is assigned directly to that DRG. Otherwise, the attribute with the next highest IG is selected from the attribute set and linked to those DRGs by an edge. It is split recursively, further on each subset of attribute values.

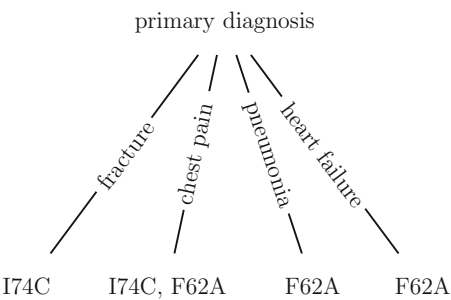
In the following, the method is illustrated based on the example in Table 2.5. For the sake of simplicity, instances 1–10 are considered. Table 2.10 provides the summary statistics of instances 1–10 for which the information entropy can be computed.

Using Eq. (2.1), the information entropy for the class attribute \mathcal{D} comes up to $H(\mathcal{D}) = 0.673$ while the conditional entropies for “gender”, “age” and “primary diagnosis” are $H(\mathcal{D}|\text{gender}) = 0.670$, $H(\mathcal{D}|\text{age}) = 0.600$ and $H(\mathcal{D}|\text{primary diagnosis}) = 0.370$, respectively. Using Eq. (2.3), the IG for the attribute “gender” is $IG(\text{gender}) = 0.003$ and for “age” and “primary diagnosis”, the information gains come up to $IG(\text{age}) = 0.073$ and $IG(\text{primary diagnosis}) = 0.303$, respectively. The results reveal that the attribute “primary diagnosis” has the highest IG and as

Table 2.10 Summary statistics of gender, age, primary diagnosis and DRG for instances 1–10 from the example provided by Table 2.5

	I74C	F62A
Gender		
Male	3	3
Female	1	3
Age		
0–30	4	2
31–100	0	4
Primary diagnosis		
Fracture	3	0
Chest pain	1	1
Pneumonia	0	4
Heart failure	0	1
DRG	4	6

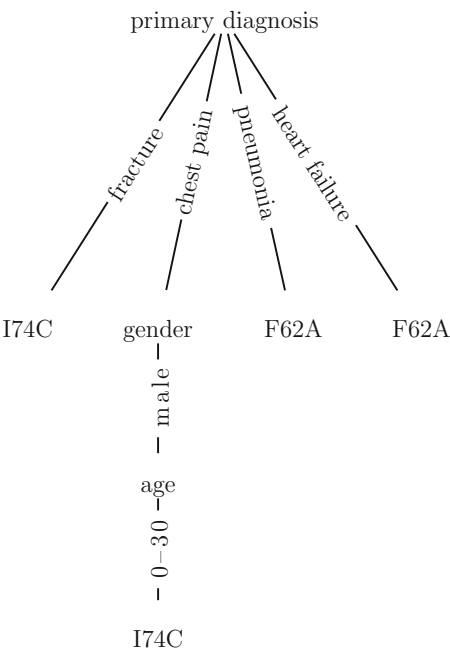
Fig. 2.6 Selected part of the classification tree before the addition of further nodes



a result, the decision tree grows by selecting this attribute as the root node, see Fig. 2.6.

Since the attribute values “fracture”, “pneumonia” and “heart failure” cannot be split further, the class is directly assigned to the respective attribute value. However, instances that contain the attribute value “chest pain” can be differentiated because here, the classes $d = I74C$ and $d = F62A$ occur. Thus, the instances are split by choosing between the remaining attributes “gender” and “age”. Both have an IG of 0 from Table 2.5 one can observe that instances in which “chest pain” occurred ($i = 4, 5$), the attribute values for “gender” and “age” are not different, conditioned on the class DRG. Therefore, we randomly assign the attribute “gender” to the attribute value “chest pain” which contains $d = I74C$ and $d = F62A$. Now, we have to split on “age”. The IG for this last attribute is, again, 0 such that we randomly select class $d = I74C$ to the attribute value “0–30” which is illustrated by Fig. 2.7. The decision tree growing process usually results in an unnecessarily large and highly specific structure. Thus, methods should be considered in order to prune the decision tree. In this dissertation, the C4.5 pruning strategy (see Witten and Frank [233]) will be evaluated. Firstly, we determine for each node the subset of training instances that is represented by the node. Secondly, we identify the DRG

Fig. 2.7 Final classification tree



that represents the majority of instances reaching the node. Thirdly, an error rate which is the number of instances not represented by this DRG is calculated. Fourthly, by specifying a confidence level (see Sect.4.1.6 for an evaluation of different confidence levels), we calculate the node’s upper error bound. Finally, we compare this bound with its children’s error rates. If the children’s combined error rates are greater than the bound, the children are pruned away from the node and replaced by a leaf.

2.3.4 Voting-Based Combined Classification

Another classification approach is to combine classifiers in order to take advantage of each individual classifier’s strengths. Different methods to combine classifiers in order to increase classification accuracy are described in Kittler et al. [103] and the following method is used in the context of this dissertation. Given the input vector of attribute values for an instance, for each DRG we count the number of classifiers which lead to the selection of this DRG. The DRG which receives the largest number of votes is then chosen while ties are resolved by employing a uniform random distribution.

2.3.5 *Probability Averaging to Combine the DRG Grouper with Machine Learning Approaches*

The following new approach has been developed in this dissertation in order to combine a DRG grouper with machine learning based classification approaches. The DRG grouper is employed as a classifier and combined with the decision tree and the Bayesian network approach. Given the probability of DRG d for classifier c as $p_{c,d}$, a new instance i is classified by employing the following rule:

$$d_i^* = \arg \max_{d \in \mathcal{D}} \frac{1}{|\mathcal{C}|} \cdot \sum_{c \in \mathcal{C}} p_{c,d}. \quad (2.12)$$

Naturally, for the two classifiers DRG grouper and decision tree, the probability for DRG d and classifier c is $p_{c,d} \in \{0, 1\}$. In contrast, in the probability distribution of the third approach, i.e. the Bayesian network classifier, the probability distribution is $p_{c,d} \in [0, 1]$. Accordingly, if the first two classifiers (DRG grouper and classification trees) support two different DRGs $d = 1$ and $d = 2$, and the Bayesian network has strong but not maximum support for a third DRG $d = 3$, i.e., $p_{3,3} \neq 100\%$ then the third DRG becomes irrelevant and the tie-breaker for choosing d_i^* is the DRG with the higher probability based on the Bayesian network's probability distribution. Then, if the probabilities for the two remaining DRGs are equal, ties are resolved by employing a uniform random distribution.

2.3.6 *Decision Rule-Based Mapping of Attribute Values to DRGs*

Holte [87] examines the performance of simple decision rules where attribute values are mapped directly to class values. For the problem of early DRG classification, the rules are determined as follows: In the training set, we count how often an attribute value of “admission diagnosis 1” occurred with respect to each DRG. For each attribute value, a mapping to the most frequent DRG is created. If two DRGs have the same frequency for the same “admission diagnosis 1”, the DRG is assigned which appeared first in the training set. The default DRG is the one with the highest frequency in the entire training set. Now, for each instance in the testing set, the value of “admission diagnosis 1” is observed and the instance to the DRG which is described by the decision rule is assigned. If the “admission diagnosis 1” has not yet been observed in the training set, the default DRG is assigned.

Optimizing Hospital-wide Patient Scheduling
Early Classification of Diagnosis-related Groups
Through Machine Learning

Gartner, D.

2014, XIV, 119 p. 22 illus., Softcover

ISBN: 978-3-319-04065-3