Chapter 6
HEAD-DT: Fitness Function Analysis

Abstract In Chap. 4, more specifically in Sect. 4.4, we saw that the definition of a fitness function for the scenario in which HEAD-DT evolves a decision-tree algorithm from multiple data sets is an interesting and relevant problem. In the experiments presented in Chap. 5, Sect. 5.2, we employed a simple average over the F-Measure obtained in the data sets that belong to the meta-training set. As previously observed, when evolving an algorithm from multiple data sets, each individual of HEAD-DT has to be executed over each data set in the meta-training set. Hence, instead of obtaining a single value of predictive performance, each individual scores a set of values that have to be eventually combined into a single measure. In this chapter, we analyse in more detail the impact of different strategies to be used as fitness function during the evolutionary cycle of HEAD-DT. We divide the experimental scheme into two distinct scenarios: (i) evolving a decision-tree induction algorithm from multiple balanced data sets; and (ii) evolving a decision-tree induction algorithm from multiple imbalanced data sets. In each of these scenarios, we analyse the difference in performance of well-known performance measures such as accuracy, F-Measure, AUC, recall, and also a lesser-known criterion, namely the relative accuracy improvement. In addition, we analyse different schemes of aggregation, such as simple average, median, and harmonic mean.

Keywords Fitness functions · Performance measures · Evaluation schemes

6.1 Performance Measures

Performance measures are key tools to assess the quality of machine learning approaches and models. Therefore, several different measures have been proposed in the specialized literature with the goal of providing better choices in general or for a specific application domain [2].

In the context of HEAD-DT's fitness function, and given that it evaluates algorithms (individuals) over data sets, it is reasonable to assume that different

Assuming we are not interested in dealing with a multi-objective optimisation problem.

R.C. Barros et al., *Automatic Design of Decision-Tree Induction Algorithms*,
SpringerBriefs in Computer Science, DOI 10.1007/978-3-319-14231-9_6

classification performance measures could be employed to provide a quantitative assessment of algorithmic performance. In the next few sections, we present five different performance measures that were selected for further investigation as HEAD-DT's fitness function.

6.1.1 Accuracy

Probably the most well-known performance evaluation measure for classification problems, the accuracy of a model is the rate of correctly classified instances:

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{6.1}$$

where $tp(tn)$ stands for the true positives (true negatives)—instances correctly classified,—and $fp(fn)$ stands for the false positives (false negatives)—instances incorrectly classified.

Even though most classification algorithms are assessed regarding the accuracy they obtain in a data set, we must point out that accuracy may be a misleading performance measure. For instance, suppose we have a data set whose class distribution is very skewed: 90 % of the instances belong to class A and 10 % to class B. An algorithm that always classifies instances as belonging to class A would achieve 90 % of accuracy, even though it never predicts a class-B instance. In this case, assuming that class B is equally important (or even more so) than class A, we would prefer an algorithm with lower accuracy, but which could eventually correctly predict some instances as belonging to the rare class B.

6.1.2 F-Measure

As it was presented in Sect. 4.4, F-Measure (also F-score or F_1 score) is the harmonic mean of precision and recall:

$$precision = \frac{tp}{tp + fp} \tag{6.2}$$

$$recall = \frac{tp}{tp + fn} \tag{6.3}$$

$$f1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{6.4}$$

Note that though F-Measure is advocated in the machine learning literature as a single measure capable of capturing the effectiveness of a system, it still completely ignores the tn, which can vary freely without affecting the statistic [8].

6.1.3 Area Under the ROC Curve

The area under the ROC (receiver operating characteristic) curve (AUC) has been increasingly used as a performance evaluation measure in classification problems. The ROC curve graphically displays the trade-off between the true positive rate ($tpr = tp/(tp + fn)$) and the false positive rate ($fpr = fp/(fp + tn)$) of a classifier. ROC graphs have properties that make them especially useful for domains with skewed class distribution and unequal classification error costs [1].

To create the ROC curve, one needs to build a graph in which the tpr is plotted along the y axis and the fpr is shown on the x axis. Each point along the curve corresponds to one of the models induced by a given algorithm, and different models are built by varying a probabilistic threshold that determines whether an instance should be classified as positive or negative.

A ROC curve is a two-dimensional depiction of a classifier. To compare classifiers, we may want to reduce ROC performance to a single scalar value representing the expected performance, which is precisely the AUC. Since the AUC is a portion of the area of the unit square, its value will always be between 0 and 1. However, because random guessing produces a diagonal line between (0,0) and (1,1), which has an area of 0.5, no realistic classifier should have an AUC value of less than 0.5. The AUC has an important statistical property: it is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance, which makes of the AUC equivalent to the Wilcoxon test of ranks [6].

The machine learning community often uses the AUC statistic for model comparison, even though this practice has recently been questioned based upon new research that shows that AUC is quite noisy as a performance measure for classification [3] and has some other significant problems in model comparison [4, 5].

6.1.4 Relative Accuracy Improvement

Originally proposed by Pappa [7], the relative accuracy improvement criterion measures the normalized improvement in accuracy of a given model over the data set's default accuracy (i.e., the accuracy achieved when using the majority class of the training data to classify the unseen data):

$$RAI_i = \begin{cases} \frac{Acc_i - DefAcc_i}{1 - DefAcc_i}, & \text{if } Acc_i > DefAcc_i \\ \frac{Acc_i - DefAcc_i}{Def\,Acc_i}, & \text{otherwise} \end{cases} \tag{6.5}$$

In Eq. (6.5), Acc_i is the accuracy achieved by a given classifier in data set i, whereas $DefAcc_i$ is the default accuracy of data set i. Note that if the improvement in accuracy is positive, i.e., the classifier accuracy is greater than the default accuracy, the improvement is normalized by the maximum possible improvement over the default accuracy ($1 - DefAcc_i$). Otherwise, the drop in the accuracy is normalized by the maximum possible drop, which is the value of the default accuracy itself. Hence, the relative accuracy improvement RAI_i regarding data set i returns a value between -1 (when $Acc_i = 0$) and 1 (when $Acc_i = 1$). Any improvement regarding the default accuracy results in a positive value, whereas any drop results in a negative value. In case $Acc_i = DefAcc_i$ (i.e., no improvement or drop in accuracy is achieved), $RAI_i = 0$, as expected.

The disadvantage of the relative accuracy improvement criterion is that it is not suitable for very imbalanced problems—data sets in which the default accuracy is really close to 1,—since high accuracy does not properly translate into high performance for these kinds of problems, as we have previously seen.

6.1.5 Recall

Also known as *sensitivity* (usually in the medical field) or *true positive rate*, recall measures the proportion of actual positives that are correctly identified as such. For instance, it may refer to the percentage of sick patients who are correctly classified as having the particular disease. In terms of the confusion matrix terms, recall is computed as follows:

$$recall = \frac{tp}{tp + fn} \tag{6.6}$$

Recall is useful for the case of imbalanced data, in which the positive class is the rare class. However, note that a classifier that always predicts the positive class will achieve a perfect recall, since recall does not take into consideration the fp values. This problem is alleviated in multi-class problems, in which each class is used in turn as the positive class, and the average of the per-class recall is taken.

6.2 Aggregation Schemes

All classification measures presented in the previous section refer to the predictive performance of a given classifier in a given data set. When evolving an algorithm from multiple data sets, HEAD-DT's fitness function is measured as the aggregated

performance of the individual in each data set that belongs to the meta-training set. We propose employing three simple strategies for combining the per-data-set performance into a single quantitative value: (i) simple average; (ii) median; and (iii) harmonic mean.

The simple average (or alternatively the arithmetic average) is computed by simply taking the average of the per-data-set values, i.e., $(1/N) \times \sum_{i=1}^{N} p_i$, for a meta-training set with N data sets and a performance measure p. It gives equal importance to the performance achieved in each data set. Moreover, it is best used in situations where there are no extreme outliers and the values are independent of each other.

The median is computed by ordering the performance values from smallest to greatest, and then taking the middle value of the ordered list. If there is an even number of data sets, since there is no single middle value, either $N/2$ or $(N/2) + 1$ can be used as middle value, or alternatively their average. The median is robust to outliers in the data (extremely large or extremely low values that may influence the simple average).

Finally, the harmonic mean is given by $\left((1/N) \times \sum_{i=1}^{N} p_i\right)^{-1}$. Unlike the simple average, the harmonic mean gives less significance to high-value outliers, providing sometimes a better picture of the average.

6.3 Experimental Evaluation

In this section, we perform an empirical evaluation of the five classification performance measures presented in Sect. 6.1 and the three aggregation schemes presented in Sect. 6.2 as fitness functions of HEAD-DT. There are a total of 15 distinct fitness functions resulting from this analysis:

1. Accuracy + Simple Average (ACC-A);
2. Accuracy + Median (ACC-M);
3. Accuracy + Harmonic Mean (ACC-H);
4. AUC + Simple Average (AUC-A);
5. AUC + Median (AUC-M);
6. AUC + Harmonic Mean (AUC-H);
7. F-Measure + Simple Average (FM-A);
8. F-Measure + Median (FM-M);
9. F-Measure + Harmonic Mean (FM-H);
10. Relative Accuracy Improvement + Simple Average (RAI-A);
11. Relative Accuracy Improvement + Median (RAI-M);
12. Relative Accuracy Improvement + Harmonic Mean (RAI-H);
13. Recall + Simple Average (TPR-A);
14. Recall + Median (TPR-M);
15. Recall + Harmonic Mean (TPR-H).

For this experiment, we employed the 67 UCI data sets described in Table 5.14 organized into two scenarios: (i) 5 balanced data sets in the meta-training set; and (ii) 5 imbalanced data sets in the training set. These scenarios were created to assess the performance of the 15 distinct fitness functions in balanced and imbalanced data, considering that some of the performance measures are explicitly designed to deal with imbalanced data whereas others are not. The term "(im)balanced" was quantitatively measured according to the imbalance ratio (IR):

$$IR = \frac{F(A_{DS})}{F(B_{DS})} \tag{6.7}$$

where $F(.)$ returns the frequency of a given class, A_{DS} is the highest-frequency class in data set DS and B_{DS} the lowest-frequency class in data set DS.

Given the size and complexity of this experiment, we did not optimise HEAD-DT's parameters as in Chap. 5, Sect. 5.2. Instead, we employed typical values found in the literature of evolutionary algorithms for decision-tree induction (the same parameters as in Chap. 5, Sect. 5.1):

- Population size: 100;
- Maximum number of generations: 100;
- Selection: tournament selection with size $t = 2$;
- Elitism rate: 5 individuals;
- Crossover: uniform crossover with 90 % probability;
- Mutation: random uniform gene mutation with 5 % probability.
- Reproduction: cloning individuals with 5 % probability.

In the next sections, we present the results for both scenarios of meta-training set. Moreover, in the end of this chapter, we perform a whole new set of experiments with the best-performing fitness functions.

6.3.1 Results for the Balanced Meta-Training Set

We randomly selected 5 balanced data sets ($IR < 1.1$) from the 67 UCI data sets described in Table 5.14 to be part of the meta-training set in this experiment: iris ($IR = 1.0$), segment ($IR = 1.0$), vowel ($IR = 1.0$), mushroom ($IR = 1.07$), and kr-vs-kp ($IR = 1.09$).

Tables 6.1 and 6.2 show the results for the 62 data sets in the meta-test set regarding accuracy and F-Measure, respectively. At the bottom of each table, the average rank is presented for the 15 versions of HEAD-DT created by varying the fitness functions. We did not present standard deviation values due to space limitations within the tables.

By careful inspection of both tables, we can see that their rankings are practically the same, with the median of the relative accuracy improvement being the

Table 6.1 Accuracy values for the 15 versions of HEAD-DT varying the fitness functions

	ACC	ACC	ACC	AUC	AUC	AUC	FM	FM	FM	RAI	RAI	RAI	TPR	TPR	TPR
	A	M	H	A	M	H	A	M	H	A	M	H	A	M	H
Abalone	0.48	0.55	0.50	0.38	0.44	0.34	0.50	0.56	0.50	0.54	0.57	0.53	0.59	0.56	0.50
Anneal	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	0.99	1.00	1.00	0.99	1.00	1.00	1.00
Arrhythmia	0.85	0.85	0.82	0.78	0.78	0.75	0.82	0.85	0.80	0.85	0.86	0.80	0.84	0.73	0.82
Audiology	0.86	0.86	0.86	0.79	0.80	0.79	0.85	0.86	0.86	0.85	0.88	0.82	0.87	0.86	0.84
Autos	0.85	0.83	0.85	0.84	0.71	0.83	0.85	0.82	0.85	0.86	0.88	0.87	0.87	0.78	0.85
Balance-scale	0.78	0.76	0.77	0.77	0.75	0.76	0.78	0.81	0.77	0.81	0.81	0.85	0.78	0.79	0.77
Breast-cancer	0.63	0.75	0.63	0.63	0.69	0.68	0.66	0.74	0.66	0.66	0.73	0.78	0.70	0.77	0.67
Breast-w	0.99	0.99	0.99	0.99	0.99	0.98	0.99	0.99	0.99	0.98	0.99	0.97	0.99	0.99	0.98
Bridges1	0.63	0.73	0.59	0.61	0.73	0.65	0.62	0.74	0.61	0.66	0.74	0.60	0.64	0.74	0.64
Bridges2	0.95	0.98	0.95	0.94	0.98	0.93	0.95	0.98	0.95	0.95	0.98	0.61	0.96	0.99	0.95
Car	0.85	0.82	0.85	0.83	0.83	0.82	0.85	0.86	0.85	0.85	0.87	0.95	0.86	0.85	0.85
cmc	0.88	0.90	0.88	0.88	0.89	0.88	0.88	0.89	0.88	0.89	0.89	0.67	0.89	0.90	0.88
Colic	0.84	0.85	0.84	0.84	0.84	0.83	0.81	0.83	0.80	0.85	0.85	0.87	0.79	0.82	0.76
Column-2C	0.89	0.87	0.89	0.87	0.87	0.87	0.89	0.89	0.89	0.89	0.89	0.88	0.90	0.87	0.89
Column-3C	0.89	0.88	0.89	0.88	0.89	0.88	0.89	0.90	0.89	0.89	0.90	0.89	0.89	0.89	0.89
Credit-a	0.77	0.82	0.82	0.77	0.79	0.73	0.79	0.83	0.80	0.77	0.84	0.89	0.81	0.80	0.71
Credit-g	0.83	0.81	0.83	0.80	0.81	0.79	0.83	0.84	0.83	0.83	0.84	0.81	0.84	0.84	0.84
Cylinder-bands	0.96	0.96	0.96	0.95	0.95	0.95	0.96	0.96	0.96	0.96	0.96	0.83	0.96	0.96	0.96
Dermatology	0.89	0.88	0.89	0.86	0.87	0.86	0.89	0.89	0.89	0.89	0.89	0.96	0.89	0.87	0.89
Diabetes	0.51	0.56	0.50	0.47	0.50	0.48	0.47	0.55	0.49	0.53	0.52	0.84	0.54	0.55	0.47
Ecoli	0.78	0.79	0.76	0.75	0.76	0.74	0.76	0.79	0.76	0.78	0.79	0.89	0.78	0.81	0.76

(continued)

Table 6.1 (continued)

	ACC A	ACC M	ACC H	AUC A	AUC M	AUC H	FM A	FM M	FM H	RAI A	RAI M	RAI H	TPR A	TPR M	TPR H
Flags	0.81	0.81	0.80	0.79	0.81	0.78	0.81	0.84	0.80	0.81	0.84	0.78	0.82	0.82	0.80
Glass	0.79	0.77	0.79	0.78	0.78	0.77	0.80	0.80	0.80	0.80	0.81	0.83	0.81	0.79	0.80
Haberman	0.86	0.86	0.86	0.87	0.87	0.86	0.86	0.87	0.86	0.87	0.87	0.79	0.86	0.87	0.86
Hayes-roth	0.85	0.84	0.85	0.84	0.84	0.84	0.85	0.87	0.85	0.86	0.87	0.86	0.86	0.87	0.85
Heart-c	0.66	0.69	0.66	0.61	0.63	0.60	0.66	0.70	0.66	0.68	0.70	0.85	0.69	0.70	0.66
Heart-h	0.94	0.94	0.94	0.93	0.93	0.93	0.94	0.94	0.94	0.94	0.94	0.84	0.95	0.94	0.94
Heart-statlog	0.87	0.84	0.89	0.87	0.85	0.86	0.88	0.87	0.86	0.85	0.89	0.85	0.85	0.85	0.84
Hepatitis	0.83	0.85	0.88	0.82	0.83	0.72	0.79	0.83	0.82	0.77	0.88	0.90	0.80	0.80	0.73
Ionosphere	0.81	0.83	0.81	0.86	0.80	0.80	0.80	0.83	0.79	0.77	0.80	0.94	0.76	0.80	0.73
kdd-synthetic	0.79	0.81	0.79	0.77	0.78	0.76	0.79	0.80	0.79	0.80	0.80	0.96	0.82	0.80	0.79
Labor	0.68	0.58	0.68	0.65	0.68	0.66	0.67	0.70	0.67	0.69	0.71	0.87	0.69	0.66	0.67
Liver-disorders	0.85	0.87	0.85	0.83	0.85	0.85	0.85	0.88	0.85	0.86	0.88	0.79	0.85	0.89	0.85
Lung-cancer	0.25	0.23	0.30	0.14	0.17	0.10	0.26	0.32	0.29	0.23	0.35	0.68	0.32	0.28	0.23
Lymph	0.77	0.79	0.78	0.75	0.77	0.75	0.78	0.79	0.78	0.78	0.79	0.85	0.79	0.79	0.78
mb-promoters	0.71	0.72	0.72	0.70	0.70	0.70	0.72	0.73	0.73	0.74	0.73	0.30	0.75	0.72	0.73
Meta.data	0.87	0.86	0.86	0.87	0.87	0.86	0.86	0.89	0.86	0.86	0.89	0.87	0.87	0.88	0.87
Morphological	0.83	0.83	0.83	0.81	0.83	0.80	0.83	0.84	0.83	0.84	0.84	0.78	0.85	0.84	0.83
Postoperative	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.71	1.00	1.00	1.00
Primary-tumor	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.51	1.00	1.00	1.00
Readings-2	0.77	0.77	0.77	0.74	0.76	0.74	0.76	0.78	0.77	0.77	0.78	1.00	0.78	0.78	0.77
Readings-4	0.77	0.77	0.77	0.76	0.77	0.76	0.77	0.79	0.77	0.78	0.79	1.00	0.78	0.79	0.77

(continued)

Table 6.1 (continued)

	ACC A	ACC M	ACC H	AUC A	AUC M	AUC H	FM A	FM M	FM H	RAI A	RAI M	RAI H	TPR A	TPR M	TPR H
Semeion	0.66	0.60	0.63	0.59	0.55	0.61	0.63	0.66	0.57	0.63	0.63	0.97	0.63	0.63	0.62
Shuttle-control	0.89	0.87	0.90	0.91	0.88	0.91	0.86	0.90	0.89	0.91	0.88	0.67	0.91	0.89	0.90
Sick	0.88	0.84	0.88	0.85	0.84	0.85	0.88	0.87	0.88	0.88	0.87	0.99	0.88	0.83	0.88
Solar-flare-1	0.95	0.94	0.95	0.94	0.93	0.93	0.95	0.96	0.95	0.96	0.96	0.77	0.95	0.95	0.95
Solar-flare-2	0.72	0.75	0.72	0.67	0.64	0.67	0.72	0.76	0.72	0.74	0.77	0.77	0.76	0.75	0.73
Sonar	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88	1.00	1.00	1.00
Soybean	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.93	0.65	0.65	0.65
Sponge	0.91	0.95	0.90	0.89	0.95	0.89	0.90	0.94	0.90	0.91	0.94	0.95	0.91	0.96	0.91
Tae	0.87	0.55	0.88	0.78	0.63	0.79	0.77	0.80	0.75	0.88	0.79	0.72	0.67	0.58	0.71
Tempdiag	0.80	0.79	0.80	0.79	0.79	0.79	0.80	0.81	0.80	0.80	0.81	1.00	0.81	0.80	0.80
Tep.fea	0.84	0.86	0.85	0.80	0.82	0.78	0.85	0.86	0.85	0.86	0.86	0.65	0.87	0.85	0.84
Tic-tac-toe	0.97	0.96	0.97	0.97	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.91	0.96	0.96	0.97
Trains	0.96	0.94	0.96	0.95	0.94	0.95	0.96	0.95	0.96	0.96	0.95	0.85	0.96	0.93	0.96
Transfusion	0.75	0.79	0.77	0.70	0.73	0.67	0.77	0.80	0.77	0.79	0.80	0.80	0.81	0.80	0.76
Vehicle	0.97	0.96	0.97	0.96	0.96	0.95	0.97	0.97	0.96	0.96	0.97	0.86	0.97	0.97	0.96
Vote	0.90	0.91	0.86	0.94	0.95	0.94	0.86	0.96	0.86	0.90	0.91	0.97	0.77	0.96	0.83
Wine	0.62	0.61	0.66	0.62	0.63	0.57	0.64	0.60	0.62	0.57	0.64	0.96	0.62	0.59	0.58
Wine-red	0.93	0.90	0.93	0.90	0.88	0.91	0.93	0.90	0.93	0.93	0.91	0.78	0.93	0.89	0.93
Wine-white	0.94	0.94	0.94	0.93	0.94	0.93	0.94	0.94	0.94	0.94	0.94	0.78	0.94	0.94	0.94
Zoo	0.59	0.61	0.59	0.56	0.58	0.54	0.59	0.62	0.59	0.61	0.61	0.90	0.61	0.61	0.58
Average rank	8.00	8.93	8.35	11.68	10.76	12.57	8.25	4.75	9.10	6.41	**3.72**	6.64	4.93	6.88	9.04

Meta-training comprises 5 balanced data sets

Table 6.2 F-Measure values for the 15 versions of HEAD-DT varying the fitness functions

	ACC	ACC	ACC	AUC	AUC	AUC	FM	FM	FM	RAI	RAI	RAI	TPR	TPR	TPR
	A	M	H	A	M	H	A	M	H	A	M	H	A	M	H
Abalone	0.48	0.55	0.50	0.37	0.43	0.33	0.50	0.56	0.50	0.53	0.57	0.53	0.59	0.56	0.49
Anneal	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	0.99	1.00	1.00	0.99	1.00	1.00	1.00
Arrhythmia	0.57	0.56	0.65	0.56	0.58	0.47	0.60	0.53	0.56	0.46	0.62	0.79	0.57	0.52	0.49
Audiology	0.84	0.83	0.80	0.75	0.75	0.73	0.81	0.84	0.78	0.84	0.85	0.81	0.83	0.68	0.81
Autos	0.85	0.86	0.86	0.79	0.79	0.78	0.85	0.86	0.86	0.85	0.88	0.87	0.87	0.86	0.84
Balance-scale	0.85	0.82	0.85	0.82	0.68	0.81	0.85	0.82	0.85	0.87	0.88	0.85	0.87	0.77	0.85
Breast-cancer	0.76	0.71	0.72	0.76	0.74	0.75	0.75	0.80	0.74	0.80	0.81	0.76	0.75	0.76	0.73
Breast-w	0.97	0.96	0.97	0.96	0.96	0.95	0.97	0.97	0.96	0.96	0.97	0.97	0.97	0.97	0.96
Bridges1	0.60	0.74	0.60	0.61	0.69	0.67	0.65	0.73	0.65	0.65	0.73	0.56	0.70	0.77	0.66
Bridges2	0.60	0.72	0.54	0.60	0.72	0.64	0.58	0.73	0.57	0.65	0.73	0.56	0.59	0.72	0.60
Car	0.95	0.98	0.95	0.94	0.98	0.93	0.95	0.98	0.95	0.95	0.98	0.95	0.96	0.99	0.95
cmc	0.66	0.69	0.65	0.61	0.63	0.60	0.66	0.69	0.66	0.68	0.70	0.66	0.69	0.70	0.66
Colic	0.82	0.85	0.88	0.81	0.82	0.67	0.75	0.82	0.81	0.73	0.88	0.87	0.77	0.77	0.68
Column-2C	0.88	0.90	0.88	0.88	0.89	0.88	0.88	0.89	0.88	0.89	0.89	0.88	0.89	0.90	0.88
Column-3C	0.89	0.87	0.89	0.87	0.87	0.87	0.89	0.89	0.89	0.89	0.89	0.89	0.90	0.86	0.89
Credit-a	0.89	0.88	0.89	0.88	0.89	0.88	0.89	0.90	0.89	0.89	0.90	0.89	0.89	0.89	0.89
Credit-g	0.80	0.78	0.80	0.78	0.81	0.78	0.80	0.84	0.80	0.81	0.84	0.80	0.82	0.79	0.80
Cylinder-bands	0.75	0.82	0.82	0.77	0.79	0.72	0.79	0.83	0.80	0.77	0.84	0.83	0.81	0.80	0.69
Dermatology	0.96	0.96	0.96	0.95	0.95	0.95	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96
Diabetes	0.83	0.82	0.83	0.81	0.83	0.80	0.83	0.84	0.83	0.84	0.84	0.84	0.85	0.84	0.83
Ecoli	0.88	0.88	0.88	0.86	0.87	0.86	0.88	0.89	0.88	0.89	0.89	0.88	0.89	0.86	0.88

(continued)

Table 6.2 (continued)

	ACC	ACC	ACC	AUC	AUC	AUC	FM	FM	FM	RAI	RAI	RAI	TPR	TPR	TPR
	A	M	H	A	M	H	A	M	H	A	M	H	A	M	H
Flags	0.78	0.79	0.76	0.74	0.75	0.74	0.76	0.79	0.76	0.78	0.79	0.78	0.78	0.81	0.76
Glass	0.83	0.81	0.83	0.79	0.80	0.78	0.83	0.83	0.83	0.83	0.84	0.83	0.84	0.84	0.84
Haberman	0.78	0.73	0.78	0.76	0.77	0.75	0.78	0.79	0.78	0.79	0.80	0.77	0.80	0.76	0.79
Hayes-roth	0.86	0.86	0.86	0.87	0.87	0.86	0.86	0.86	0.86	0.86	0.87	0.86	0.86	0.86	0.86
Heart-c	0.85	0.82	0.85	0.83	0.83	0.82	0.85	0.86	0.85	0.85	0.87	0.85	0.86	0.85	0.85
Heart-h	0.84	0.85	0.84	0.84	0.84	0.83	0.79	0.81	0.78	0.85	0.85	0.83	0.75	0.80	0.72
Heart-statlog	0.85	0.84	0.85	0.84	0.84	0.84	0.85	0.87	0.85	0.86	0.87	0.85	0.86	0.87	0.85
Hepatitis	0.85	0.80	0.89	0.86	0.83	0.83	0.86	0.86	0.84	0.82	0.88	0.89	0.83	0.83	0.82
Ionosphere	0.94	0.94	0.94	0.93	0.93	0.93	0.94	0.94	0.94	0.94	0.94	0.94	0.95	0.94	0.94
kdd-synthetic	0.93	0.90	0.93	0.90	0.88	0.91	0.93	0.90	0.93	0.93	0.91	0.96	0.93	0.88	0.93
Labor	0.78	0.80	0.79	0.85	0.78	0.77	0.78	0.83	0.76	0.72	0.76	0.87	0.71	0.77	0.67
Liver-disorders	0.79	0.81	0.79	0.77	0.78	0.76	0.79	0.80	0.79	0.80	0.80	0.79	0.82	0.80	0.79
Lung-cancer	0.68	0.54	0.68	0.65	0.68	0.66	0.67	0.70	0.67	0.69	0.71	0.68	0.69	0.64	0.67
Lymph	0.85	0.87	0.85	0.83	0.84	0.85	0.85	0.88	0.85	0.86	0.88	0.85	0.85	0.89	0.85
mb-promoters	0.87	0.86	0.86	0.86	0.87	0.86	0.86	0.89	0.86	0.86	0.89	0.28	0.87	0.88	0.86
Meta.data	0.24	0.21	0.28	0.12	0.16	0.08	0.25	0.32	0.28	0.24	0.35	0.86	0.32	0.28	0.23
Morphological	0.77	0.79	0.77	0.75	0.76	0.74	0.77	0.78	0.77	0.78	0.78	0.78	0.79	0.78	0.78
Postoperative	0.65	0.66	0.65	0.68	0.66	0.68	0.68	0.72	0.68	0.72	0.72	0.65	0.70	0.69	0.66
Primary-tumor	0.48	0.54	0.47	0.43	0.46	0.43	0.44	0.53	0.45	0.50	0.50	0.48	0.51	0.54	0.43
Readings-2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Readings-4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Semeion	0.94	0.94	0.94	0.93	0.94	0.93	0.94	0.94	0.94	0.94	0.94	0.97	0.94	0.94	0.94

(continued)

Table 6.2 (continued)

	ACC A	ACC M	ACC H	AUC A	AUC M	AUC H	FM A	FM M	FM H	RAI A	RAI M	RAI H	TPR A	TPR M	TPR H
Shuttle-control	0.63	0.53	0.57	0.52	0.49	0.53	0.58	0.63	0.48	0.58	0.60	0.63	0.56	0.56	0.55
Sick	0.99	0.99	0.99	0.99	0.99	0.98	0.99	0.99	0.99	0.98	0.99	0.99	0.99	0.99	0.98
Solar-flare-1	0.77	0.76	0.76	0.73	0.75	0.73	0.76	0.77	0.76	0.77	0.77	0.76	0.77	0.77	0.76
Solar-flare-2	0.77	0.76	0.76	0.75	0.76	0.75	0.76	0.78	0.76	0.78	0.78	0.77	0.77	0.78	0.76
Sonar	0.88	0.84	0.88	0.85	0.84	0.85	0.88	0.87	0.88	0.88	0.87	0.88	0.88	0.83	0.88
Soybean	0.88	0.85	0.90	0.90	0.87	0.91	0.85	0.89	0.88	0.90	0.87	0.93	0.90	0.87	0.89
Sponge	0.95	0.92	0.94	0.94	0.91	0.92	0.94	0.96	0.94	0.96	0.96	0.95	0.94	0.94	0.93
Tae	0.72	0.75	0.72	0.67	0.64	0.67	0.72	0.76	0.72	0.74	0.77	0.72	0.76	0.75	0.73
Tempdiag	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Tep.fea	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61
Tic-tac-toe	0.91	0.95	0.90	0.89	0.95	0.89	0.91	0.94	0.90	0.91	0.94	0.91	0.91	0.96	0.91
Trains	0.87	0.53	0.88	0.78	0.61	0.79	0.77	0.80	0.75	0.88	0.79	0.85	0.66	0.56	0.71
Transfusion	0.79	0.75	0.79	0.78	0.78	0.78	0.79	0.80	0.79	0.80	0.80	0.79	0.80	0.77	0.79
Vehicle	0.84	0.86	0.85	0.80	0.82	0.78	0.85	0.86	0.85	0.86	0.86	0.86	0.87	0.85	0.85
Vote	0.97	0.96	0.97	0.97	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.96	0.96	0.97
Wine	0.96	0.94	0.96	0.95	0.94	0.95	0.96	0.95	0.96	0.96	0.95	0.96	0.96	0.93	0.96
Wine-red	0.75	0.79	0.77	0.69	0.73	0.66	0.77	0.80	0.77	0.79	0.80	0.78	0.81	0.80	0.76
Wine-white	0.58	0.61	0.59	0.55	0.58	0.53	0.59	0.62	0.59	0.61	0.61	0.78	0.61	0.62	0.58
Zoo	0.90	0.90	0.84	0.94	0.95	0.94	0.84	0.96	0.84	0.90	0.91	0.90	0.75	0.96	0.81
Average rank	7.94	9.22	8.45	11.30	10.56	12.35	8.17	4.61	9.16	6.27	**3.60**	6.64	5.25	7.17	9.31

Meta-training comprises 5 balanced data sets

Table 6.3 Values are the average performance (rank) of each version of HEAD-DT according to either accuracy or F-Measure

Version	Accuracy rank	F-Measure rank	Average
ACC-A	8.00	7.94	7.97
ACC-M	8.93	9.22	9.08
ACC-H	8.35	8.45	8.40
AUC-A	11.68	11.30	11.49
AUC-M	10.76	10.56	10.66
AUC-H	12.57	12.35	12.46
FM-A	8.25	8.17	8.21
FM-M	4.75	4.61	4.68
FM-H	9.10	9.16	9.13
RAI-A	6.41	6.27	6.34
RAI-M	**3.72**	**3.60**	**3.66**
RAI-H	6.64	6.64	6.64
TPR-A	4.93	5.25	5.09
TPR-M	6.88	7.17	7.03
TPR-H	9.04	9.31	9.18

best-ranked method for either evaluation measure. Only a small position-switching occurs between the accuracy and F-Measure rankings, with respect to the positions of ACC-M, TPR-H, and FM-H.

Table 6.3 summarizes the average rank values obtained by each version of HEAD-DT with respect to accuracy and F-Measure. Values in bold indicate the best performing version according to the corresponding evaluation measure. It can be seen that version RAI-M is the best-performing method regardless of the evaluation measure. The average of the average ranks (average across evaluation measures) indicates the following final ranking positions (from best to worst): (1) RAI-M; (2) FM-M; (3) TPR-A; (4) RAI-A; (5) RAI-H; (6) TPR-M; (7) ACC-A; (8) FM-A; (9) ACC-H; (10) ACC-M; (11) FM-H; (12) TPR-H; (13) AUC-M; (14) AUC-A; (15) AUC-H.

For evaluating whether the differences between versions are statistically significant, we present the critical diagrams of the accuracy and F-Measure values in Fig. 6.1. It is possible to observe that there are no significant differences among the top-4 versions (RAI-M, FM-M, TPR-A, and RAI-A). Nevertheless, RAI-M is the only version that outperforms TPR-M and RAI-H with statistical significance in both evaluation measures, which is not the case of FM-M, TPR-A, and RAI-A.

Some interesting conclusions can be drawn from this first set of experiments with a balanced meta-training set:

- The AUC measure was not particularly effective for evolving decision-tree algorithms in this scenario, regardless of the aggregation scheme being used. Note that versions of HEAD-DT that employ AUC in their fitness function perform quite

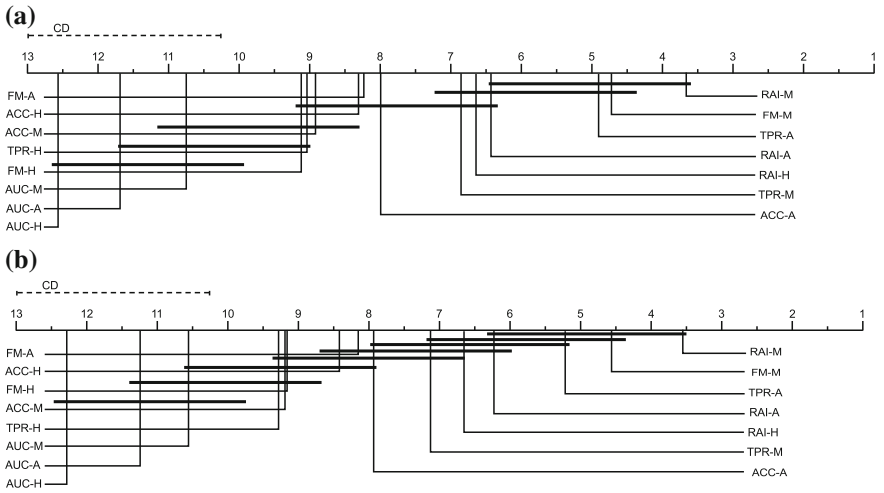

Fig. 6.1 Critical diagrams for the balanced meta-training set experiment. **a** Accuracy rank. **b** F-measure rank

poorly when compared to the remaining versions—AUC-M, AUC-A, and AUC-H are in the bottom of the ranking: 13th, 14th, and 15th position, respectively;

- The use of the harmonic mean as an aggregation scheme was not successful overall. The harmonic mean was often worst aggregation scheme for the evaluation measures, occupying the lower positions of the ranking (except when combined to RAI).
- The use of the median, on the other hand, was shown to be very effective in most cases. For 3 evaluation measures the median was the best aggregation scheme (relative accuracy improvement, F-Measure, and AUC). In addition, the two best-ranked versions made use of the median as their aggregation scheme;
- The relative accuracy improvement was overall the best evaluation measure, occupying the top part of the ranking (1st, 4th, and 5th best-ranked versions);
- Finally, both F-Measure and recall were consistently among the best versions (2nd, 3rd, 6th, and 8th best-ranked versions), except once again when associated to the harmonic mean (11th and 12th).

Figure 6.2 depicts a picture of the fitness evolution throughout the evolutionary cycle. It presents both the best fitness from the population at a given generation and the average fitness from the corresponding generation.

Note that version AUC-M (Fig. 6.2e) achieves the perfect fitness from the very first generation (AUC = 1). We further analysed this particular case and verified that the decision-tree algorithm designed in this version does not perform any kind of pruning. Even though prune-free algorithms usually overfit the training data (if no pre-pruning is performed as well, they achieve 100 % of accuracy in the training data) and thus underperform in the test data, it seems that this was not the case for the 5 data sets in the meta-training set. In the particular validation sets of the meta-training

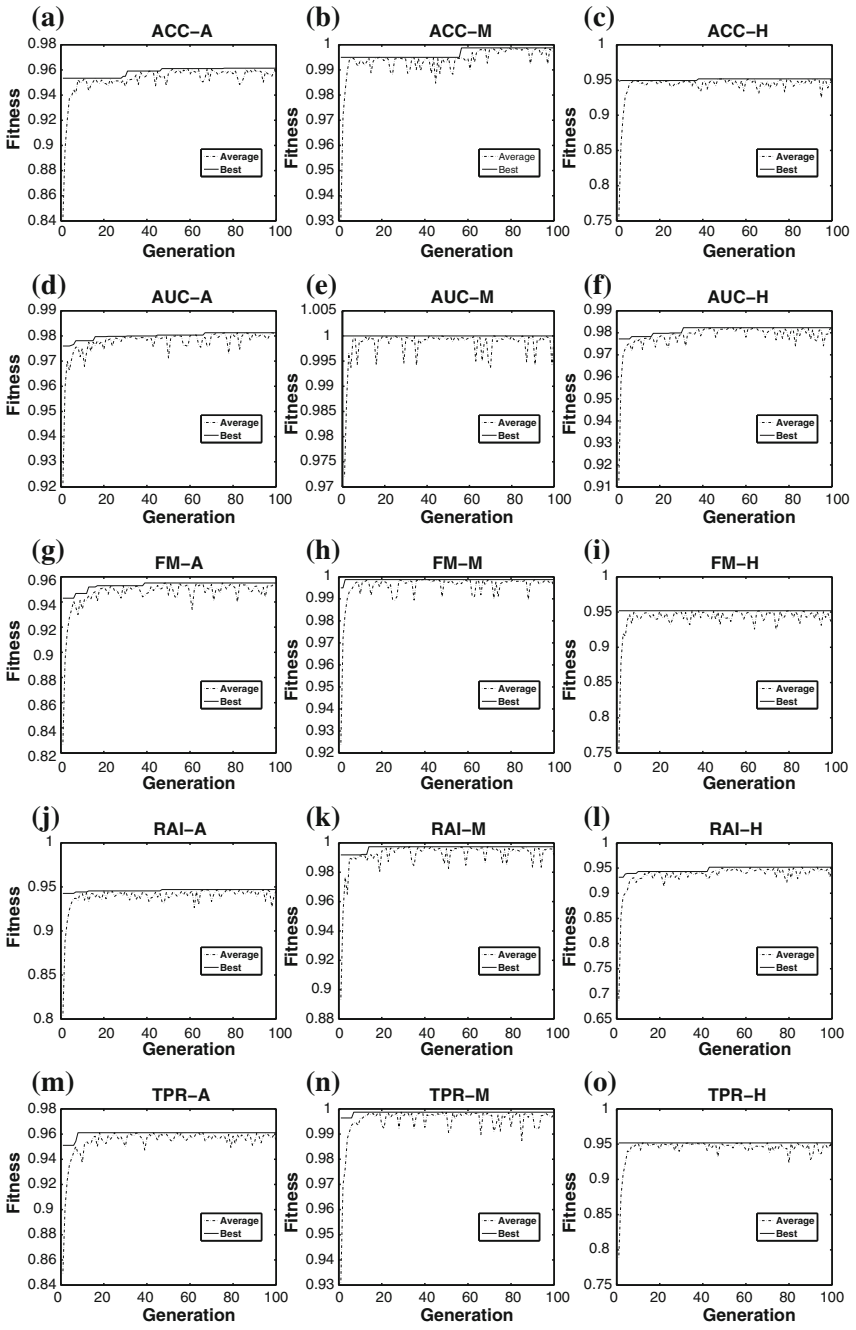

Fig. 6.2 Fitness evolution in HEAD-DT for the balanced meta-training set

set, a prune-free algorithm with the stop criterion *minimum number of 3 instances* was capable of achieving perfect AUC. Nevertheless, this automatically-designed algorithm certainly suffered from overfitting in the meta-test set, since AUC-M was only the 13th-best out of 15 versions.

Versions FM-H (Fig. 6.2i) and TPR-H (Fig. 6.2o) also achieved their best fitness value in the first generation. The harmonic mean, due to its own nature (ignore higher values), seems to make the search for better individuals harder than the other aggregation schemes.

6.3.2 Results for the Imbalanced Meta-Training Set

We randomly selected 5 imbalanced data sets ($IR > 10$) from the 67 UCI data sets described in Table 5.14 to be part of the meta-training set in this experiment: primary-tumor ($IR = 84$), anneal ($IR = 85.5$), arrhythmia ($IR = 122.5$), winequality-white ($IR = 439.6$), and abalone ($IR = 689$).

Tables 6.4 and 6.5 show the results for the 62 data sets in the meta-test set regarding accuracy and F-Measure, respectively. At the bottom of each table, the average rank is presented for the 15 versions of HEAD-DT created by varying the fitness functions. We once again did not present standard deviation values due to space limitations within the tables.

By careful inspection of both tables, we can see that the rankings in them are practically the same, with the average F-Measure being the best-ranked method for either evaluation measure. Only a small position-switching occurs between the accuracy and F-Measure rankings, with respect to the positions of ACC-H and RAI-M.

Table 6.6 summarizes the average rank values obtained by each version of HEAD-DT with respect to accuracy and F-Measure. Values in bold indicate the best performing version according to the corresponding evaluation measure. Note that version FM-A is the best-performing method regardless of the evaluation measure. The average of the average ranks (average across evaluation measures) indicates the following final ranking positions (from best to worst): (1) FM-A; (2) TPR-A; (3) TPR-H; (4) AUC-A; (5) AUC-H; (6) FM-H; (7) ACC-A; (8) ACC-M; (9) ACC-H; (10) RAI-M; (11) RAI-H; (12) FM-M; (13) TPR-M; (14) RAI-A; (15) AUC-M.

For evaluating whether the differences among the versions are statistically significant, we present the critical diagrams of the accuracy and F-Measure values in Fig. 6.3. We can see that there are no statistically significant differences among the 7 (5) best-ranked versions regarding accuracy (F-Measure). In addition, note that the 6 best-ranked versions involve performance measures that are suitable for evaluating imbalanced problems (F-Measure, recall, and AUC), which is actually expected given the composition of the meta-training set.

The following conclusions can be drawn from this second set of experiments concerning imbalanced data sets:

Table 6.4 Accuracy values for the 15 versions of HEAD-DT varying the fitness functions

	ACC	ACC	ACC	AUC	AUC	AUC	FM	FM	FM	RAI	RAI	RAI	TPR	TPR	TPR
	A	M	H	A	M	H	A	M	H	A	M	H	A	M	H
Audiology	0.67	0.65	0.69	0.75	0.60	0.60	0.59	0.61	0.60	0.59	0.59	0.55	0.60	0.59	0.60
Autos	0.79	0.74	0.72	0.84	0.63	0.76	0.74	0.49	0.55	0.44	0.47	0.71	0.77	0.35	0.69
Balance-scale	0.82	0.79	0.81	0.80	0.66	0.69	0.72	0.71	0.71	0.58	0.71	0.80	0.71	0.71	0.71
Breast-cancer	0.74	0.73	0.72	0.74	0.60	0.86	0.86	0.70	0.84	0.68	0.84	0.72	0.85	0.62	0.86
Bridges1	0.67	0.69	0.66	0.77	0.53	0.87	0.89	0.82	0.84	0.72	0.78	0.68	0.84	0.75	0.84
Bridges2	0.64	0.67	0.66	0.73	0.53	0.68	0.70	0.58	0.69	0.55	0.60	0.68	0.67	0.56	0.66
Car	0.94	0.92	0.94	0.91	0.76	1.00	1.00	0.94	1.00	0.80	1.00	0.93	1.00	0.94	1.00
Heart-c	0.80	0.80	0.80	0.83	0.67	0.76	0.79	0.75	0.76	0.64	0.76	0.80	0.78	0.74	0.79
cmc	0.60	0.58	0.57	0.57	0.50	0.73	0.74	0.60	0.67	0.59	0.66	0.57	0.74	0.59	0.73
Column-2C	0.85	0.84	0.83	0.86	0.69	0.68	0.74	0.61	0.67	0.60	0.66	0.83	0.73	0.56	0.74
Column-3C	0.84	0.84	0.82	0.87	0.70	0.66	0.71	0.50	0.57	0.54	0.59	0.83	0.68	0.47	0.65
Credit-a	0.87	0.88	0.87	0.88	0.70	0.93	0.95	0.83	0.94	0.75	0.91	0.87	0.95	0.77	0.95
Cylinder-bands	0.78	0.74	0.72	0.78	0.62	0.77	0.81	0.72	0.76	0.64	0.77	0.73	0.80	0.68	0.82
Dermatology	0.96	0.95	0.93	0.96	0.73	0.86	0.88	0.84	0.87	0.69	0.84	0.92	0.88	0.84	0.88
Ecoli	0.84	0.85	0.84	0.86	0.68	0.87	0.88	0.85	0.85	0.69	0.84	0.84	0.88	0.84	0.88
Flags	0.72	0.68	0.64	0.71	0.56	0.96	0.95	0.92	0.94	0.75	0.91	0.66	0.95	0.92	0.94
Credit-g	0.76	0.75	0.75	0.77	0.63	0.74	0.79	0.70	0.78	0.62	0.73	0.75	0.78	0.70	0.77
Glass	0.79	0.79	0.75	0.78	0.62	0.84	0.87	0.77	0.84	0.68	0.80	0.76	0.86	0.71	0.85
Haberman	0.77	0.75	0.75	0.76	0.62	0.85	0.86	0.82	0.86	0.69	0.85	0.76	0.87	0.83	0.87
Hayes-roth	0.85	0.84	0.78	0.85	0.60	0.73	0.77	0.73	0.73	0.61	0.74	0.81	0.76	0.72	0.75
Heart-statlog	0.82	0.81	0.80	0.82	0.67	0.93	0.94	0.92	0.95	0.76	0.94	0.80	0.94	0.92	0.93
Hepatitis	0.81	0.81	0.81	0.89	0.68	0.89	0.86	0.81	0.82	0.66	0.80	0.80	0.86	0.81	0.85

(continued)

Table 6.4 (continued)

	ACC A	ACC M	ACC H	AUC A	AUC M	AUC H	FM A	FM M	FM H	RAI A	RAI M	RAI H	TPR A	TPR M	TPR H
Colic	0.87	0.85	0.85	0.86	0.67	0.79	0.84	0.79	0.84	0.67	0.80	0.86	0.82	0.79	0.82
Heart-h	0.82	0.81	0.81	0.82	0.66	0.82	0.85	0.82	0.82	0.68	0.82	0.81	0.84	0.83	0.85
Ionosphere	0.93	0.91	0.91	0.92	0.73	0.79	0.79	0.78	0.78	0.62	0.77	0.92	0.81	0.78	0.80
Iris	0.95	0.95	0.95	0.96	0.77	0.82	0.84	0.81	0.82	0.67	0.81	0.95	0.83	0.80	0.84
kr-vs-kp	0.96	0.96	0.96	0.96	0.78	0.73	0.74	0.72	0.72	0.59	0.73	0.96	0.75	0.72	0.74
Labor	0.85	0.77	0.83	0.86	0.68	0.83	0.83	0.79	0.80	0.65	0.80	0.78	0.83	0.79	0.83
Liver-disorders	0.71	0.74	0.71	0.75	0.61	0.83	0.85	0.70	0.83	0.66	0.73	0.73	0.84	0.69	0.80
Lung-cancer	0.62	0.65	0.55	0.65	0.47	0.72	0.74	0.62	0.74	0.58	0.66	0.64	0.74	0.60	0.73
Lymph	0.84	0.80	0.80	0.85	0.65	0.85	0.89	0.78	0.83	0.70	0.83	0.79	0.87	0.67	0.89
Meta.data	0.13	0.12	0.11	0.10	0.13	0.78	0.83	0.77	0.81	0.66	0.79	0.11	0.82	0.78	0.82
Morphological	0.74	0.73	0.73	0.73	0.60	0.95	0.96	0.95	0.96	0.77	0.94	0.73	0.97	0.94	0.96
mb-promoters	0.86	0.85	0.80	0.88	0.63	0.96	0.96	0.95	0.96	0.77	0.95	0.85	0.96	0.96	0.96
Mushroom	0.99	0.98	0.99	0.99	0.80	0.86	0.88	0.85	0.86	0.70	0.86	0.99	0.87	0.82	0.86
Diabetes	0.81	0.78	0.78	0.78	0.64	0.87	0.95	0.86	0.95	0.75	0.87	0.78	0.95	0.80	0.95
Postoperative	0.72	0.71	0.71	0.70	0.57	0.86	0.85	0.79	0.78	0.66	0.78	0.71	0.86	0.77	0.83
Segment	0.96	0.94	0.95	0.94	0.77	0.86	0.90	0.86	0.88	0.71	0.88	0.95	0.90	0.79	0.89
Semeion	0.94	0.92	0.93	0.94	0.76	0.10	0.22	0.09	0.13	0.17	0.13	0.93	0.22	0.08	0.18
Readings-2	0.94	0.98	1.00	1.00	0.80	0.92	0.94	0.90	0.93	0.75	0.92	0.95	0.94	0.88	0.94
Readings-4	0.94	0.98	1.00	1.00	0.80	0.92	0.97	0.91	0.96	0.77	0.95	0.95	0.96	0.79	0.97
Shuttle-control	0.60	0.61	0.61	0.57	0.61	0.87	0.98	0.92	0.97	0.78	0.93	0.60	0.98	0.84	0.98
Sick	0.97	0.95	0.98	0.98	0.79	0.75	0.79	0.69	0.70	0.58	0.59	0.97	0.81	0.59	0.78

(continued)

Table 6.4 (continued)

	ACC A	ACC M	ACC H	AUC A	AUC M	AUC H	FM A	FM M	FM H	RAI A	RAI M	RAI H	TPR A	TPR M	TPR H
Solar-flare-1	0.72	0.73	0.73	0.74	0.60	0.75	0.77	0.75	0.75	0.61	0.75	0.73	0.76	0.74	0.76
Solar-flare2	0.76	0.75	0.76	0.75	0.61	0.72	0.76	0.72	0.75	0.60	0.72	0.75	0.77	0.72	0.76
Sonar	0.80	0.80	0.79	0.84	0.67	1.00	0.95	1.00	0.94	0.75	0.98	0.80	1.00	1.00	1.00
Soybean	0.79	0.88	0.83	0.87	0.67	0.58	0.63	0.57	0.62	0.49	0.58	0.68	0.62	0.57	0.62
Sponge	0.93	0.93	0.92	0.95	0.74	0.77	0.83	0.75	0.81	0.65	0.74	0.92	0.83	0.75	0.82
kdd-synthetic	0.92	0.92	0.91	0.95	0.74	0.62	0.74	0.62	0.71	0.57	0.64	0.91	0.75	0.60	0.74
Tae	0.66	0.62	0.59	0.70	0.53	0.77	0.83	0.74	0.81	0.65	0.72	0.61	0.82	0.69	0.81
Tempdiag	1.00	1.00	0.91	1.00	0.77	0.77	0.81	0.75	0.78	0.64	0.76	0.95	0.81	0.75	0.81
Tep.fea	0.65	0.65	0.65	0.65	0.52	0.81	0.80	0.79	0.82	0.63	0.81	0.65	0.93	0.58	0.90
Tic-tac-toe	0.90	0.88	0.88	0.90	0.73	1.00	0.95	1.00	0.94	0.75	0.98	0.90	1.00	1.00	1.00
Trains	0.59	0.48	0.37	0.75	0.39	0.65	0.65	0.65	0.65	0.52	0.65	0.51	0.65	0.65	0.65
Transfusion	0.77	0.77	0.79	0.79	0.64	0.95	0.94	0.90	0.93	0.74	0.92	0.78	0.95	0.90	0.94
Vehicle	0.79	0.76	0.74	0.76	0.64	0.94	0.96	0.93	0.95	0.76	0.93	0.74	0.96	0.94	0.96
Vote	0.96	0.96	0.96	0.96	0.77	0.98	0.98	0.97	0.98	0.78	0.96	0.96	0.99	0.98	0.98
Vowel	0.73	0.75	0.74	0.67	0.62	0.96	0.96	0.95	0.96	0.76	0.95	0.70	0.95	0.95	0.95
Wine	0.93	0.90	0.90	0.96	0.75	0.94	0.99	0.95	0.98	0.78	0.94	0.91	0.99	0.91	0.99
Wine-red	0.69	0.65	0.63	0.62	0.55	0.99	1.00	0.99	1.00	0.80	0.98	0.63	1.00	0.98	1.00
Breast-w	0.95	0.94	0.95	0.95	0.76	0.94	0.97	0.95	0.97	0.78	0.95	0.95	0.98	0.95	0.97
Zoo	0.94	0.92	0.89	0.93	0.65	0.66	0.82	0.65	0.79	0.61	0.70	0.92	0.88	0.64	0.87
Average rank	6.70	7.94	8.40	5.87	13.43	6.70	**4.02**	9.71	6.70	13.40	8.58	8.72	4.19	10.53	5.10

Meta-training comprises 5 imbalanced data sets

Table 6.5 F-Measure values for the 15 versions of HEAD-DT varying the fitness functions

	ACC	ACC	ACC	AUC	AUC	AUC	FM	FM	FM	RAI	RAI	RAI	TPR	TPR	TPR
	A	M	H	A	M	H	A	M	H	A	M	H	A	M	H
Audiology	0.63	0.60	0.64	0.71	0.56	0.58	0.54	0.51	0.52	0.52	0.50	0.48	0.48	0.46	0.47
Autos	0.79	0.74	0.71	0.84	0.63	0.76	0.73	0.46	0.54	0.43	0.45	0.71	0.76	0.30	0.68
Balance-scale	0.81	0.76	0.77	0.78	0.65	0.66	0.67	0.59	0.63	0.53	0.61	0.77	0.60	0.59	0.61
Breast-cancer	0.68	0.67	0.68	0.73	0.59	0.86	0.86	0.67	0.84	0.68	0.84	0.65	0.85	0.62	0.86
Bridges1	0.62	0.63	0.60	0.77	0.51	0.87	0.89	0.81	0.82	0.71	0.75	0.63	0.83	0.70	0.83
Bridges2	0.57	0.60	0.59	0.72	0.50	0.68	0.70	0.58	0.69	0.55	0.60	0.61	0.67	0.56	0.66
Car	0.93	0.92	0.94	0.91	0.76	1.00	1.00	0.94	1.00	0.80	1.00	0.93	1.00	0.94	1.00
Heart-c	0.80	0.80	0.80	0.83	0.67	0.74	0.78	0.69	0.70	0.62	0.71	0.80	0.76	0.69	0.77
cmc	0.59	0.57	0.57	0.57	0.49	0.73	0.73	0.53	0.61	0.58	0.59	0.57	0.72	0.53	0.71
Column-2C	0.85	0.84	0.83	0.86	0.69	0.67	0.73	0.52	0.62	0.58	0.59	0.83	0.71	0.48	0.72
Column-3C	0.83	0.84	0.82	0.87	0.70	0.66	0.71	0.40	0.53	0.54	0.57	0.82	0.68	0.40	0.65
Credit-a	0.87	0.88	0.87	0.88	0.70	0.93	0.95	0.78	0.94	0.75	0.89	0.87	0.95	0.73	0.95
Cylinder-bands	0.78	0.73	0.72	0.78	0.62	0.77	0.81	0.70	0.75	0.63	0.76	0.73	0.80	0.67	0.81
Dermatology	0.96	0.95	0.93	0.96	0.73	0.86	0.88	0.84	0.87	0.69	0.84	0.91	0.88	0.84	0.87
Ecoli	0.83	0.84	0.83	0.85	0.68	0.87	0.88	0.85	0.84	0.69	0.84	0.83	0.88	0.83	0.88
Flags	0.70	0.67	0.62	0.71	0.56	0.96	0.95	0.92	0.94	0.75	0.91	0.63	0.95	0.92	0.94
Credit-g	0.72	0.74	0.72	0.76	0.62	0.74	0.79	0.69	0.78	0.62	0.72	0.73	0.78	0.69	0.77
Glass	0.78	0.78	0.74	0.77	0.61	0.84	0.87	0.75	0.84	0.68	0.79	0.74	0.85	0.70	0.85
Haberman	0.71	0.71	0.70	0.75	0.61	0.84	0.86	0.80	0.86	0.68	0.84	0.70	0.86	0.81	0.86
Hayes-roth	0.85	0.83	0.78	0.85	0.59	0.72	0.76	0.68	0.68	0.60	0.70	0.81	0.74	0.68	0.73
Heart-statlog	0.82	0.81	0.80	0.82	0.67	0.93	0.93	0.88	0.94	0.75	0.92	0.80	0.93	0.89	0.89
Hepatitis	0.76	0.75	0.76	0.88	0.66	0.88	0.85	0.76	0.77	0.64	0.76	0.74	0.84	0.76	0.83

(continued)

Table 6.5 (continued)

	ACC A	ACC M	ACC H	AUC A	AUC M	AUC H	FM A	FM M	FM H	RAI A	RAI M	RAI H	TPR A	TPR M	TPR H
Colic	0.87	0.85	0.85	0.86	0.66	0.76	0.83	0.77	0.83	0.67	0.77	0.86	0.80	0.77	0.80
Heart-h	0.82	0.80	0.80	0.82	0.66	0.82	0.85	0.82	0.82	0.68	0.82	0.81	0.84	0.83	0.84
Ionosphere	0.93	0.91	0.91	0.92	0.73	0.77	0.78	0.73	0.74	0.62	0.73	0.92	0.78	0.76	0.78
Iris	0.95	0.95	0.95	0.96	0.77	0.82	0.84	0.80	0.82	0.67	0.81	0.95	0.83	0.80	0.84
kr-vs-kp	0.96	0.96	0.96	0.96	0.78	0.72	0.73	0.69	0.70	0.58	0.70	0.96	0.73	0.70	0.72
Labor	0.83	0.72	0.83	0.85	0.68	0.83	0.83	0.79	0.80	0.65	0.79	0.74	0.83	0.79	0.83
Liver-disorders	0.69	0.73	0.70	0.75	0.61	0.83	0.85	0.69	0.83	0.66	0.73	0.72	0.84	0.68	0.80
Lung-cancer	0.60	0.64	0.48	0.65	0.42	0.71	0.74	0.59	0.73	0.57	0.64	0.63	0.73	0.56	0.72
Lymph	0.83	0.80	0.78	0.85	0.65	0.85	0.89	0.77	0.83	0.70	0.83	0.78	0.87	0.67	0.89
Meta.data	0.11	0.10	0.09	0.09	0.12	0.78	0.83	0.77	0.81	0.65	0.78	0.08	0.82	0.77	0.82
Morphological	0.72	0.71	0.72	0.72	0.60	0.95	0.96	0.95	0.96	0.77	0.94	0.72	0.97	0.94	0.96
mb-promoters	0.86	0.84	0.80	0.88	0.63	0.96	0.96	0.95	0.96	0.77	0.95	0.85	0.96	0.96	0.96
Mushroom	0.99	0.98	0.99	0.99	0.80	0.86	0.88	0.85	0.86	0.70	0.86	0.99	0.87	0.82	0.86
Diabetes	0.80	0.78	0.77	0.78	0.64	0.87	0.95	0.86	0.94	0.75	0.86	0.77	0.95	0.80	0.95
Postoperative	0.63	0.59	0.59	0.65	0.51	0.86	0.85	0.79	0.78	0.66	0.78	0.59	0.86	0.77	0.83
Segment	0.96	0.94	0.95	0.94	0.77	0.85	0.90	0.86	0.88	0.71	0.88	0.95	0.90	0.78	0.89
Semeion	0.94	0.90	0.92	0.93	0.76	0.09	0.21	0.07	0.11	0.16	0.11	0.92	0.21	0.05	0.17
Readings-2	0.93	0.98	1.00	1.00	0.80	0.92	0.94	0.89	0.93	0.75	0.92	0.94	0.94	0.88	0.94
Readings-4	0.93	0.98	1.00	1.00	0.80	0.92	0.97	0.91	0.96	0.77	0.95	0.94	0.96	0.79	0.97
Shuttle-control	0.52	0.49	0.47	0.55	0.52	0.87	0.98	0.92	0.97	0.78	0.92	0.47	0.98	0.83	0.98
Sick	0.97	0.94	0.98	0.98	0.79	0.71	0.77	0.63	0.68	0.56	0.54	0.96	0.79	0.51	0.75

(continued)

Table 6.5 (continued)

	ACC	ACC	ACC	AUC	AUC	AUC	FM	FM	FM	RAI	RAI	RAI	TPR	TPR	TPR
	A	M	H	A	M	H	A	M	H	A	M	H	A	M	H
Solar-flare-1	0.71	0.71	0.71	0.72	0.59	0.74	0.76	0.73	0.73	0.60	0.73	0.70	0.75	0.72	0.74
Solar-flare2	0.74	0.73	0.73	0.74	0.60	0.72	0.76	0.70	0.74	0.60	0.72	0.73	0.77	0.69	0.76
Sonar	0.80	0.80	0.79	0.84	0.67	1.00	0.93	1.00	0.93	0.74	0.97	0.80	1.00	1.00	1.00
Soybean	0.77	0.87	0.81	0.86	0.66	0.57	0.63	0.57	0.62	0.49	0.58	0.65	0.62	0.57	0.61
Sponge	0.91	0.91	0.88	0.94	0.73	0.76	0.83	0.74	0.81	0.65	0.74	0.88	0.83	0.74	0.82
kdd-synthetic	0.92	0.92	0.91	0.95	0.74	0.61	0.74	0.60	0.71	0.56	0.63	0.91	0.75	0.57	0.73
Tae	0.66	0.61	0.59	0.70	0.53	0.76	0.83	0.73	0.81	0.65	0.72	0.61	0.82	0.69	0.80
Tempdiag	1.00	1.00	0.91	1.00	0.77	0.76	0.81	0.71	0.75	0.64	0.75	0.95	0.81	0.73	0.80
Tep.fea	0.61	0.61	0.60	0.61	0.49	0.79	0.78	0.77	0.80	0.61	0.78	0.61	0.93	0.51	0.89
Tic-tac-toe	0.90	0.88	0.88	0.91	0.73	1.00	0.93	1.00	0.93	0.74	0.97	0.90	1.00	1.00	1.00
Trains	0.59	0.47	0.33	0.75	0.37	0.61	0.61	0.61	0.61	0.49	0.61	0.49	0.61	0.61	0.61
Transfusion	0.73	0.71	0.77	0.77	0.63	0.96	0.94	0.90	0.93	0.74	0.92	0.73	0.95	0.90	0.94
Vehicle	0.79	0.75	0.74	0.76	0.64	0.93	0.96	0.92	0.95	0.76	0.92	0.73	0.96	0.94	0.96
Vote	0.96	0.96	0.96	0.96	0.77	0.98	0.98	0.97	0.98	0.78	0.96	0.96	0.99	0.98	0.98
Vowel	0.73	0.75	0.74	0.66	0.62	0.96	0.96	0.95	0.96	0.76	0.95	0.69	0.95	0.95	0.95
Wine	0.93	0.90	0.90	0.96	0.75	0.94	0.99	0.95	0.98	0.78	0.94	0.91	0.99	0.91	0.99
Wine-red	0.68	0.63	0.61	0.61	0.54	0.99	1.00	0.99	1.00	0.80	0.98	0.61	1.00	0.98	1.00
Breast-w	0.95	0.94	0.95	0.95	0.76	0.94	0.97	0.95	0.97	0.78	0.95	0.95	0.98	0.95	0.97
Zoo	0.94	0.91	0.86	0.93	0.62	0.65	0.82	0.65	0.79	0.61	0.69	0.91	0.88	0.64	0.87
Average rank	6.92	8.23	8.74	5.44	13.23	6.25	**3.83**	9.97	6.79	12.94	8.65	8.95	4.27	10.56	5.25

Meta-training comprises 5 imbalanced data sets

Table 6.6 Values are the average performance (rank) of each version of HEAD-DT according to either accuracy or F-Measure

Version	Accuracy rank	F-Measure rank	Average
ACC-A	6.70	6.92	6.81
ACC-M	7.94	8.23	8.09
ACC-H	8.40	8.74	8.57
AUC-A	5.87	5.44	5.66
AUC-M	13.43	13.23	13.33
AUC-H	6.70	6.25	6.48
FM-A	**4.02**	**3.83**	**3.93**
FM-M	9.71	9.97	9.84
FM-H	6.70	6.79	6.75
RAI-A	13.40	12.94	13.17
RAI-M	8.58	8.65	8.62
RAI-H	8.72	8.95	8.84
TPR-A	4.19	4.27	4.23
TPR-M	10.53	10.56	10.55
TPR-H	5.10	5.25	5.18

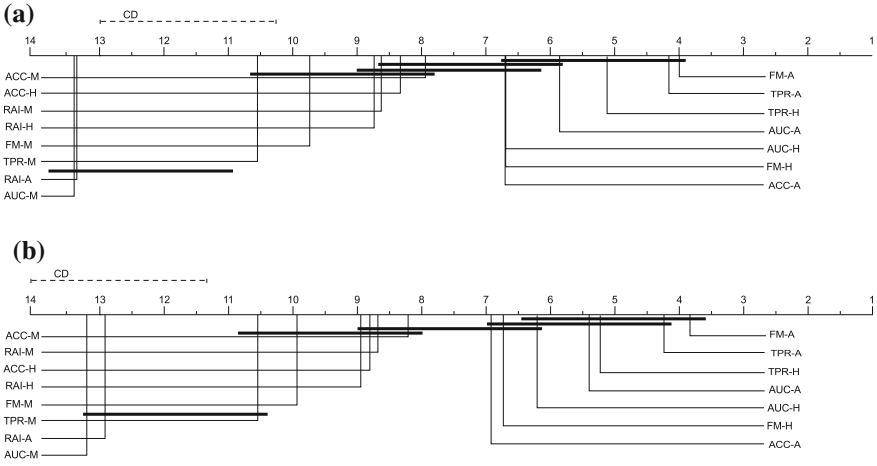

Fig. 6.3 Critical diagrams for the imbalanced meta-training set experiment. **a** Accuracy rank. **b** F-measure rank

- The relative accuracy improvement is not suitable for dealing with imbalanced data sets and hence occupies the bottom positions of the ranking (10th, 11th, and 14th positions). This behavior is expected given that RAI measures the improvement over the majority-class accuracy, and such an improvement is often damaging for imbalanced problems, in which the goal is to improve the accuracy of the less-frequent class(es);
- The median was the worst aggregation scheme overall, figuring in the bottom positions of the ranking (8th, 10th, 12th, 13th, and 15th). It is interesting to notice that the median was very successful in the balanced meta-training experiment, and quite the opposite in the imbalanced one;
- The simple average, on the other hand, presented itself as the best aggregation scheme for the imbalanced data, figuring in the top of the ranking (1st, 2nd, 4th, 7th), except when associated to RAI (14th), which was the worst performance measure overall;
- The 6 best-ranked versions were those employing performance measures known to be suitable for imbalanced data (F-Measure, recall, and AUC);
- Finally, the harmonic mean had a solid performance throughout this experiment, differently from its performance in the balanced meta-training experiment.

Figure 6.4 depicts a picture of the fitness evolution throughout the evolutionary cycle. Note that whereas some versions find their best individual at the very end of evolution (e.g., FM-H, Fig. 6.4i), others converge quite early (e.g., TPR-H, Fig. 6.4o), though there seems to exist no direct relation between early (or late) convergence and predictive performance.

6.3.3 Experiments with the Best-Performing Strategy

Considering that the median of the relative accuracy improvement (RAI-M) was the best-ranked fitness function for the balanced meta-training set, and that the average F-Measure (FM-A) was the best-ranked fitness function for the imbalanced meta-training set, we perform a comparison of these HEAD-DT versions with the baseline decision-tree induction algorithms C4.5, CART, and REPTree.

For version RAI-M, we use the same meta-training set as before: iris ($IR = 1$), segment ($IR = 1$), vowel ($IR = 1$), mushroom ($IR = 1.07$), and kr-vs-kp ($IR = 1.09$). The resulting algorithm is tested over the 10 most-balanced data sets from Table 5.14:

1. meta-data ($IR = 1$);
2. mfeat ($IR = 1$);
3. mb-promoters ($IR = 1$);
4. kdd-synthetic ($IR = 1$);
5. trains ($IR = 1$);
6. tae ($IR = 1.06$);
7. vehicle ($IR = 1.10$);

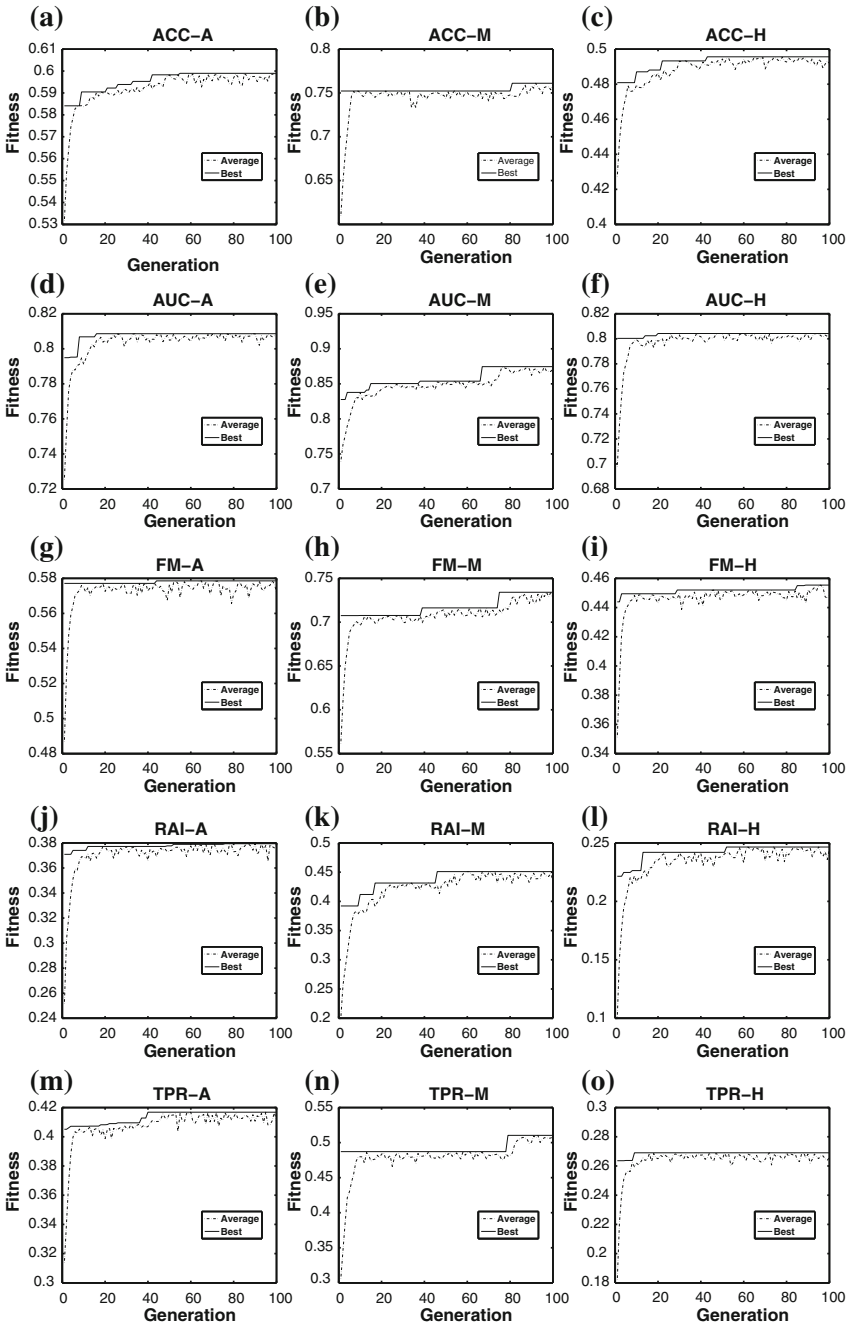

Fig. 6.4 Fitness evolution in HEAD-DT for the imbalanced meta-training set

8. sonar ($IR = 1.14$);
9. heart-c ($IR = 1.20$);
10. credit-a ($IR = 1.25$).

For version FM-A, we also use the same meta-training set as before: primary-tumor ($IR = 84$), anneal ($IR = 85.5$), arrhythmia ($IR = 122.5$), winequality-white ($IR = 439.6$), and abalone ($IR = 689$). The resulting algorithm is tested over the 10 most-imbalanced data sets from Table 5.14:

- flags ($IR = 15$);
- sick ($IR = 15.33$);
- car ($IR = 18.62$);
- autos ($IR = 22.33$);
- sponge ($IR = 23.33$);
- postoperative ($IR = 32$);
- lymph ($IR = 40.50$);
- audiology ($IR = 57$);
- winequality-red ($IR = 68.10$);
- ecoli ($IR = 71.50$).

In Chap. 5, we saw that HEAD-DT is capable of generating effective algorithms tailored to a particular application domain (gene expression data). Now, with this new experiment, our goal is to verify whether HEAD-DT is capable of generating effective algorithms tailored to a particular statistical profile—in this case, tailored to balanced/imbalanced data.

Table 6.7 shows the accuracy and F-Measure values for HEAD-DT, C4.5, CART, and REPTree, in the 20 UCI data sets (10 most-balanced and 10 most-imbalanced). The version of HEAD-DT executed over the first 10 data sets is RAI-M, whereas the version executed over the remaining 10 is FM-A. In both versions, HEAD-DT is executed 5 times as usual, and the results are averaged.

Observe in Table 6.7 that HEAD-DT (RAI-M) outperforms C4.5, CART, and REPTree in 8 out of 10 data sets (in both accuracy and F-Measure), whereas C4.5 is the best algorithm in the remaining two data sets. The same can be said about HEAD-DT (FM-A), which also outperforms C4.5, CART, and REPTree in 8 out of 10 data sets, being outperformed once by C4.5 and once by CART.

We proceed by presenting the critical diagrams of accuracy and F-Measure (Fig. 6.5) in order to evaluate whether the differences among the algorithms are statistically significant. Note that HEAD-DT is the best-ranked method, often in the 1st position (rank = 1.30). HEAD-DT (versions RAI-M and FM-A) outperforms both CART and REPTree with statistical significance for $\alpha = 0.05$. With respect to C4.5, it is outperformed by HEAD-DT with statistical significance for $\alpha = 0.10$, though not for $\alpha = 0.05$. Nevertheless, we are confident that being the best method in 16 out of 20 data sets is enough to conclude that HEAD-DT automatically generates decision-tree algorithms tailored to balanced/imbalanced data that are consistently more effective than C4.5, CART, and REPTree.

Table 6.7 Accuracy and F-Measure values for the 10 most-balanced data sets and the 10 most-imbalanced data sets

Data set	IR	HEAD-DT		C4.5		CART		REP	
		Accuracy	F-Measure	Accuracy	F-Measure	Accuracy	F-Measure	Accuracy	F-Measure
Meta.data	1.00	**0.35 ± 0.09**	**0.35 ± 0.10**	0.04 ± 0.03	0.02 ± 0.02	0.05 ± 0.03	0.02 ± 0.01	0.04 ± 0.00	0.00 ± 0.00
mfeat	1.00	**0.79 ± 0.01**	**0.78 ± 0.02**	0.72 ± 0.02	0.70 ± 0.02	0.72 ± 0.04	0.70 ± 0.04	0.72 ± 0.03	0.70 ± 0.03
mb-promoters	1.00	**0.89 ± 0.03**	**0.89 ± 0.03**	0.80 ± 0.13	0.79 ± 0.14	0.72 ± 0.14	0.71 ± 0.14	0.77 ± 0.15	0.76 ± 0.15
kdd-synthetic	1.00	0.91 ± 0.03	0.91 ± 0.03	**0.91 ± 0.04**	**0.91 ± 0.04**	0.88 ± 0.04	0.88 ± 0.04	0.88 ± 0.03	0.87 ± 0.04
Trains	1.00	0.79 ± 0.06	0.79 ± 0.06	**0.90 ± 0.32**	**0.90 ± 0.32**	0.20 ± 0.42	0.20 ± 0.42	0.00 ± 0.00	0.00 ± 0.00
Tae	1.06	**0.77 ± 0.03**	**0.77 ± 0.03**	0.60 ± 0.11	0.59 ± 0.12	0.51 ± 0.12	0.49 ± 0.15	0.47 ± 0.12	0.45 ± 0.12
Vehicle	1.10	**0.86 ± 0.01**	**0.86 ± 0.01**	0.74 ± 0.04	0.73 ± 0.04	0.72 ± 0.04	0.71 ± 0.05	0.71 ± 0.04	0.70 ± 0.04
Sonar	1.14	**0.87 ± 0.01**	**0.87 ± 0.01**	0.73 ± 0.08	0.72 ± 0.08	0.71 ± 0.06	0.71 ± 0.06	0.71 ± 0.07	0.70 ± 0.07
Heart-c	1.20	**0.87 ± 0.01**	**0.87 ± 0.01**	0.77 ± 0.09	0.76 ± 0.09	0.81 ± 0.04	0.80 ± 0.04	0.77 ± 0.08	0.77 ± 0.08
Credit-a	1.25	**0.90 ± 0.01**	**0.90 ± 0.01**	0.85 ± 0.03	0.85 ± 0.03	0.84 ± 0.03	0.84 ± 0.03	0.85 ± 0.03	0.85 ± 0.03
Flags	15.00	**0.74 ± 0.01**	**0.74 ± 0.01**	0.63 ± 0.05	0.61 ± 0.05	0.61 ± 0.10	0.57 ± 0.10	0.62 ± 0.10	0.58 ± 0.10
Sick	15.33	0.98 ± 0.01	0.98 ± 0.01	0.99 ± 0.00	0.99 ± 0.00	**0.99 ± 0.01**	**0.99 ± 0.01**	0.99 ± 0.01	0.99 ± 0.01
Car	18.62	**0.98 ± 0.01**	**0.98 ± 0.01**	0.93 ± 0.02	0.93 ± 0.02	0.97 ± 0.02	0.97 ± 0.02	0.89 ± 0.02	0.89 ± 0.02
Autos	22.33	0.85 ± 0.03	0.85 ± 0.03	**0.86 ± 0.06**	**0.85 ± 0.07**	0.78 ± 0.10	0.77 ± 0.10	0.65 ± 0.08	0.62 ± 0.07
Sponge	23.33	**0.94 ± 0.01**	**0.93 ± 0.02**	0.93 ± 0.06	0.89 ± 0.09	0.91 ± 0.06	0.88 ± 0.09	0.91 ± 0.08	0.88 ± 0.10
Postoperative	32.00	**0.72 ± 0.01**	**0.67 ± 0.04**	0.70 ± 0.05	0.59 ± 0.07	0.71 ± 0.06	0.59 ± 0.08	0.69 ± 0.09	0.58 ± 0.09
Lymph	40.50	**0.87 ± 0.01**	**0.87 ± 0.01**	0.78 ± 0.09	0.79 ± 0.10	0.75 ± 0.12	0.73 ± 0.14	0.77 ± 0.11	0.76 ± 0.12
Audiology	57.00	**0.79 ± 0.04**	**0.77 ± 0.05**	0.78 ± 0.07	0.75 ± 0.08	0.74 ± 0.05	0.71 ± 0.05	0.74 ± 0.08	0.70 ± 0.09
Wine-red	68.10	**0.74 ± 0.02**	**0.74 ± 0.02**	0.61 ± 0.03	0.61 ± 0.03	0.63 ± 0.02	0.61 ± 0.02	0.60 ± 0.03	0.58 ± 0.03
Ecoli	71.50	**0.86 ± 0.01**	**0.86 ± 0.01**	0.84 ± 0.07	0.83 ± 0.07	0.84 ± 0.07	0.82 ± 0.07	0.79 ± 0.09	0.77 ± 0.09
	Rank	**1.30**	**1.30**	2.25	2.25	2.90	2.90	3.55	3.55

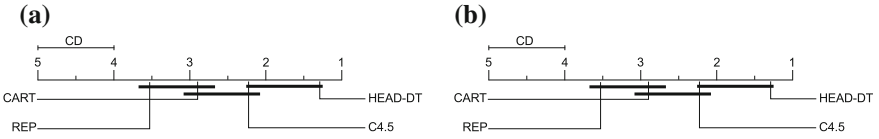

Fig. 6.5 Critical diagrams for accuracy and F-Measure. Values are regarding the 20 UCI data sets in Table 6.7. **a** Accuracy rank for the balanced data sets. **b** F-measure rank for the balanced data sets

Since HEAD-DT is run 5 times for alleviating the randomness effect of evolutionary algorithms, we further analyse the 5 algorithms generated by HEAD-DT for the balanced meta-training set and the 5 algorithms generated for the imbalanced meta-training set.

Regarding the balanced meta-training set, we noticed that the favored split criterion was the G statistic (present in 40 % of the algorithms). The favored stop criterion was stopping the tree-splitting process only when there is a single instance in the node (present in 80 % of the algorithms). The homogeneous stop was present in the remaining 20 % of the algorithms, but since a single instance is always homogeneous (only 1 class represented in the node), we can say that HEAD-DT stop criterion was actually stop splitting nodes when they are homogeneous. Surprisingly, the favored pruning strategy was not to use any pruning strategy (80 % of the algorithms). It seems that this particular combination of design components did not lead to overfitting, even though the trees were not pruned at any point. Algorithm 1 shows this custom algorithm designed for balanced data sets.

Algorithm 1 Custom algorithm designed by HEAD-DT (RAI-M) for balanced data sets.

1: Recursively split nodes using the *G* statistic;
2: Perform nominal splits in multiple subsets;
3: Perform step 1 until class-homogeneity;
4: Do not perform any pruning strategy;
When dealing with missing values:
5: Calculate the split of missing values by weighting the split criterion value;
6: Distribute missing values by weighting them according to partition probability;
7: For classifying an instance with missing values, halt in the current node.

Regarding the imbalanced meta-training set, we noticed that two split criteria stand out: DCSM (present in 40 % of the algorithms) and Normalized Gain (also present in 40 % of the algorithms). In 100 % of the algorithms, the nominal splits were aggregated into binary splits. The favored stop criterion was either the homogeneous stop (60 % of the algorithms) of the algorithms or tree stop when a maximum depth of around 10 levels is reached (40 % of the algorithms). Finally, the pruning strategy was also divided between PEP pruning with 1 SE (40 % of the algorithms) and no pruning at all (40 % of the algorithms). We noticed that whenever the algorithm employed DCSM, PEP pruning was the favored pruning strategy. Similarly, whenever the Normalized Gain was selected, *no pruning* was the favored pruning strategy. It

seems that HEAD-DT was capable of detecting a correlation between different split criteria and pruning strategies. Algorithm 2 shows the custom algorithm that was tailored to imbalanced data (we actually present the choices of different components when it was the case).

Algorithm 2 Custom algorithm designed by HEAD-DT (FM-A) for imbalanced data sets.

1: Recursively split nodes using either DCSM or the Normalized Gain;
2: Aggregate nominal splits into binary subsets;
3: Perform step 1 until class-homogeneity or a maximum depth of 9 (10) levels;
4: Either do not perform pruning and remove nodes that do not reduce training error, or perform PEP pruning with 1 SE;
When dealing with missing values:
5: Ignore missing values or perform unsupervised imputation when calculating the split criterion;
6: Perform unsupervised imputation before distributing missing values;
7: For classifying an instance with missing values, halt in the current node or explore all branches and combine the classification.

Regarding the missing value strategies, we did not notice any particular pattern in either the balanced or the imbalanced scenarios. Hence, the missing-value strategies presented in Algorithms 1 and 2 are only examples of selected components, though they did not stand out in terms of appearance frequency.

6.4 Chapter Remarks

In this chapter, we performed a series of experiments to analyse in more detail the impact of different fitness functions during the evolutionary cycle of HEAD-DT. In the first part of the chapter, we presented 5 classification performance measures and three aggregation schemes to combine these measures during fitness evaluation of multiple data sets. The combination of performance measures and aggregation schemes resulted in 15 different versions of HEAD-DT.

We designed two experimental scenarios to evaluate the 15 versions of HEAD-DT. In the first scenario, HEAD-DT is executed on a meta-training set with 5 balanced data sets, and on a meta-test set with the remaining 62 available UCI data sets. In the second scenario, HEAD-DT is executed on a meta-training set with 5 imbalanced data sets, and the meta-test set with the remaining 62 available UCI data sets. For measuring the level of data set balance, we make use of the imbalance ratio (*IR*), which is the ratio between the most-frequent and the less-frequent classes of the data.

Results of the experiments indicated that the median of the relative accuracy improvement was the most suitable fitness function for the balanced scenario, whereas the average of the F-Measure was the most suitable fitness function for the imbalanced scenario. The next step of the empirical analysis was to compare these versions of HEAD-DT with the baseline decision-tree induction algorithms C4.5, CART, and REPTree. For such, we employed the same meta-training sets than before, though the meta-test sets exclusively comprised balanced (imbalanced) data

sets. The experimental results confirmed that HEAD-DT can generate algorithms tailored to a particular statistical profile (data set balance) that are more effective than C4.5, CART, and REPTree, outperforming them in 16 out of 20 data sets.

References

1. T. Fawcett, An introduction to ROC analysis. Pattern Recognit. Lett. **27**(8), 861–874 (2006)
2. C. Ferri, J. Hernández-Orallo, R. Modroiu, An experimental comparison of performance measures for classification. Pattern Recognit. Lett. **30**(1), 27–38 (2009)
3. B. Hanczar et al., Small-sample precision of ROC-related estimates. Bioinformatics **26**(6), 822–830 (2010)
4. D.J. Hand, Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach. Learn. **77**(1), 103–123 (2009)
5. J.M. Lobo, A. Jiménez-Valverde, R. Real, AUC: a misleading measure of the performance of predictive distribution models. Glob. Ecol. Biogeogr. **17**(2), 145–151 (2008)
6. S.J. Mason, N.E. Graham, Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: statistical significance and interpretation. Q. J. R. Meteorol. Soc. **128**(584), 2145–2166 (2002)
7. G.L. Pappa, Automatically evolving rule induction algorithms with grammar-based genetic programming, Ph.D. thesis. University of Kent at Canterbury (2007)
8. D. Powers, Evaluation: From precision, recall and f-measure to ROC, informedness, markedness and correlation. J. Mach. Learn. Technol. **2**(1), 37–63 (2011)