2003－ＩＣＳ－133 (25)
2003／9／15

Experimental Evaluation of Time-series Decision Tree

Yuu YAMADA[†], Einoshin SUZUKI[†], Hideto YOKOI[††], and Katsuhiko TAKABAYASHI[††]

† Faculty of Engineering, Yokohama National University　79–5 Tokiwadai, Hodogaya, Yokohama, 240–8501 Japan

†† Division of Medical Informatics, Chiba University Hospital　1-8-1 Inohana, Chuo-ku, Chiba, 260–8677 Japan

E-mail: †yuu@slab.dnj.ynu.ac.jp,suzuki@ynu.ac.jp,
††yokoih@telemed.ho.chiba-u.ac.jp,takaba@ho.chiba-u.ac.jp

Abstract　In this paper, we give experimental evaluation of our time-series decision tree induction method under various conditions. It has been empirically observed that the method induces accurate and comprehensive decision trees in time-series classification, which has gaining increasing attention due to its importance in various real-world applications. The evaluation has revealed several important findings including interaction between a split test and its goodness.

Key words　Time-series Decision Tree, Time-series Classification, Split Test, Misclassification cost, Medical Test Data

時系列決定木の実験的評価

山田　悠[†]　鈴木英之進[†]　横井　英人[††]　高林克日己[††]

† 横浜国立大学 大学院工学研究院　〒240–8501 横浜市保土ヶ谷区常盤台 79–5

†† 千葉大学 医学部附属病院医療情報部　〒260–8677 千葉市中央区亥鼻 1-8-1

E-mail: †yuu@slab.dnj.ynu.ac.jp,suzuki@ynu.ac.jp,
††yokoih@telemed.ho.chiba-u.ac.jp,takaba@ho.chiba-u.ac.jp

あらまし　本論文では，われわれが提案した時系列決定木を種々の条件下で実験的に評価する．われわれの手法は，実応用における重要性ゆえに注目を集めつつある時系列分類学習において，正確で可読性が高い決定木を学習することが経験的に分かっている．本評価により，分割テストとその評価基準間の相互影響など，重要な知見が得られた．

キーワード　時系列決定木，時系列分類学習，分割テスト，誤分類コスト，医療検査データ

1. Introduction

Time-series data are employed in various domains including politics, economics, science, industry, agriculture, and medicine. Classification of time-series data is related to many promising application problems. For instance, an accurate classifier for liver cirrhosis from time-series data of medical tests might replace a biopsy which picks liver tissue by inserting an instrument directly into liver. Such a classifier is highly important since it would substantially reduce costs of both patients and hospital.

Our time-series decision tree represents a novel classifier for time-series classification. Our learning method for the time-series decision tree has enabled us to discover a classifier which is highly appraised by domain experts [8]. In this paper, we perform additional experiments based on advice from domain experts, and investigate on various characteristics of our time-series decision tree.

2. Time-series Decision Tree

2.1 Time-series Classification

A time sequence $\boldsymbol{A}$ represents a list of values $\alpha_1, \alpha_2, \cdots, \alpha_I$ sorted in chronological order. For simplicity, this paper assumes that the values are obtained or sampled with an equivalent interval $(= 1)$.

A data set D consists of n examples e_1, e_2, $\cdots$, e_n, and each example e_i is described by m attributes a_1, a_2, $\cdots$, a_m and a class attribute c. An attribute a_j can represent a time-series attribute which takes a time sequence as its value. The class attribute c represents a nominal attribute and its value is called a class. In time-series classification, the objective represents induction of a classifier, which predicts the class of an example e, given a training data set D.

2.2 Learning Time-series Decision Tree

Our time-series tree [8] has a time sequence which exists in data and an attribute in its internal node, and splits a set of examples according to the dissimilarity of their corresponding time sequences to the time sequence. The use of a time sequence which exists in data in its split node contributes to comprehensibility of the classifier, and each time sequence is obtained by exhaustive search. The dissimilarity measure is based on dynamic time warping (DTW) [6].

We call this split test a standard-example split test. A standard-example split test $\sigma(e, a, \theta)$ consists of a standard example e, an attribute a, and a threshold θ. Let a value of an example e in terms of a time-series attribute a be $e(a)$, then a standard-example split test divides a set of examples $e_1, e_2, \cdots, e_n$ to a set $S_1(e, a, \theta)$ of examples each of which $e_i(a)$ satisfies $G(e(a), e_i(a)) < \theta$ and the rest $S_2(e, a, \theta)$. We also call this split test a θ-guillotine cut.

As the goodness of a split test, we have selected gain ratio [7] since it is frequently used in decision-tree induction. Since at most $n - 1$ split points are inspected for an attribute in a θ-guillotine cut and we consider each example as a candidate of a standard example, it frequently happens that several split points exhibit the largest value of gain ratio. We assume that consideration on shapes of time sequences is essential in comprehensibility of a classifier, thus, in such a case, we define that the best split test exhibits the largest gap between the sets of time sequences in the child nodes. The gap $gap(e, a, \theta)$ of $\sigma(e, a, \theta)$ is equivalent to $G(e'(a), e''(a))$ where e' and e'' represent the example $e_i(a)$ in $S_1(e, a, \theta)$ with the largest $G(e(a), e_i(a))$ and the example $e_j(a)$ in $S_2(e, a, \theta)$ with the smallest $G(e(a), e_j(a))$ respectively. When several split tests exhibit the largest value of gain ratio, the split test with the largest $gap(e, a, \theta)$ among them is selected.

We have also proposed a cluster-example split test $\sigma'(e', e'', a)$ for comparison. A cluster-example split test divides a set of examples $e_1, e_2, \cdots, e_n$ into a set $U_1(e', e'', a)$ of examples each of which $e_i(a)$ satisfies $d(e'(a), e_i(a)) < d(e''(a), e_i(a))$ and the rest $U_2(e', e'', a)$. The goodness of a split test is equivalent to that of the standard-example split test without θ.

2.3 Experimental Results and Comments from Domain Experts

We have evaluated our method with Chronic hepatitis data [1], the Australian sign language data [4], and the EEG data [4]. As a result of pre-processing, we have obtained two data sets, which we call H1 and H2, from Chronic hepatitis data. Similarly, two data sets, which we call Sign and EEG, have been generated from the Australian sign language data and the EEG data respectively. The classification tasks in H1 and H2 are prediction of liver cirrhosis from medical tests data. We have employed time sequences each of which has more than 9 test values during a period of before 500 days and after 500 days of a biopsy. In both data sets, there are 30 examples of liver cirrhosis and 34 examples of the other class. Since the intervals of medical tests differ, we have employed liner interpolation between two adjacent values and transformed each time sequence to a time sequence of 101 values with a 10-day interval. In H1, one of us, who is a physician, suggested to use in classification 14 attributes (GOT, GPT, ZTT, TTT, T-BIL, I-BIL, D-BIL, T-CHO, TP, ALB, CHE, WBC, PLT, HGB) which are important in hepatitis. In H2, we have measured shifts for each of these attributes from its average value and employed these 14 attributes in addition to the original attributes. Experimental results have confirmed that our induction method constructs comprehensive and accurate decision trees.

We have prepared another data set, which we call H0, from the chronic hepatitis data by dealing with the first biopsies only. H0 consists of 51 examples (21 LC patients, and 30 non-LC patients) each of which is described with 14 attributes. Experimental results show that our time-series tree has been shown to be promising for knowledge discovery. We show a time-series decision tree learned from H0 in figure 1.

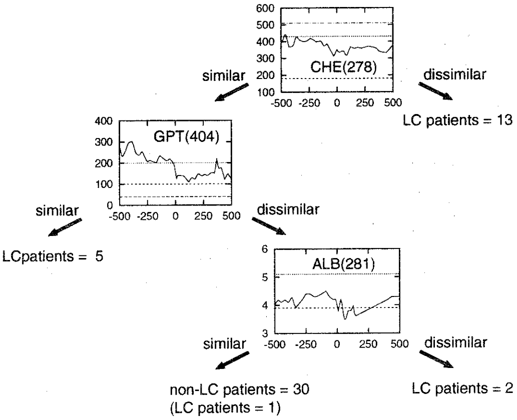

Figure 1 Time-series tree learned from H0 (chronic hepatitis data of the first biopsies)

We obtained the following comments from medical experts.

- The proposed learning method exhibits novelty and is highly interesting. The splits in the upper parts of the time-decision trees are valid, and the learning results are surprisingly well as a method which employs little domain knowledge.

- Medical test values which are measured after a biopsy are typically influenced by treatment such as interferon (IFN). It would be better to use only medical test values which were measured before a biopsy.
- 1000 days are long as a period of measurement since the number n of patients is small. It would be better to use shorter periods such as 365 days.
- The number of medical tests might be possibly reduced to 4 per year. Prediction from a few number of medical tests has a higher impact on clinical treatment.
- A medical expert is familiar with sensitivity, specificity, and an ROC curve as evaluation indices of a classifier. It causes more problems to overlook an LC patient than mistake a non-LC patient.

3. Experiments for Misclassification Costs

3.1 Conditions of Experiments

Based on a comment in the previous section, we evaluated our time-series decision tree without using medical test data after a biopsy. For a continuous attribute, C4.5 [7] employs a split test which verifies whether a value is greater than a threshold. This split test will be called an average-split test in this paper. We call our approach which employs both the standard-example split test and the average-split test combined-split test. For sake of comparison, we also employed the average-split test alone and Line-split test, which replaces a standard example by a line segment. A line segment in the latter method is obtained by discretizing test values by an equal-frequency method with $\alpha - 1$ bins, and connecting two points (-500, p_1) and (0, p_2) where each of p_1 and p_2 represents one of the end values of discretized bins. For instance, it considers 25 line segments if $\alpha = 5$. The cluster-example split test was not employed since it exhibited poor performance in [8].

Table 1 Confusion matrix

	LC	non-LC
LC (Prediction)	TP	FP
non-LC (Prediction)	FN	TN

We show a confusion matrix in table 1. As the domain experts stated, it is important to decrease the number FN of overlooked LC patients than the number FP of mistaken non-LC patients. Therefore, we employ sensitivity, specificity, and (misclassification) cost in addition to predictive accuracy as evaluation indices. The added indices are considered to be important in the following order.

$$Cost = \frac{C\,FN + FP}{C(TP + FN) + (TN + FP)} \quad (1)$$

$$Sensitivity \text{ (True Positive Rate)} = \frac{TP}{TP + FN} \quad (2)$$

$$Specificity \text{ (True Negative Rate)} = \frac{TN}{TN + FP} \quad (3)$$

where C represents a user-specified weight. We settled $C = 5$ throughout the experiments, and employed a leave-one-out method. Note that $Cost$ is normalized in order to facilitate comparison of experimental results from different data sets.

It is reported that Laplace correction is effective in decision tree induction for cost-sensitive classification [2]. We obtained the probability $\Pr(a)$ of a class a when there are $\nu(a)$ examples of a among ν examples as follows.

$$\Pr(a) = \frac{\nu(a) + l}{\nu + 2l} \quad (4)$$

where l represents a parameter of the Laplace correction. We settled $l = 1$ unless stated.

We modified data selection criteria in each series of experiments and prepared various data sets as shown in table 2. In a name of a data set, the first figure represents the selected period of measurement before a biopsy, the figure subsequent to a "p" represents the number of required medical tests, and the figure subsequent to an "i" represents the number of days of an interval in interpolation. Since we employed both B-type patients and C-type patients in all experiments, each name of a data set contains strings "BC". Since we had obtained novel data of biopsies after [8], we employed an integrated version in the experiments.

3.2 Experimental Results

Firstly, we modified the required number of medical tests to 6, 3, 2 under a 180-day period and a 5-day interpolation interval. We show the results in table 3. From the table, we see that the average-split test and the line-split test outperform other methods in cost for p2 and p6 respectively. For p3, the methods exhibit the same cost and outperform our standard-example split test. We believe that the poor performance of our method is due to lack of information on shapes of time sequences and the number of examples. We interpret the results that lack of the former information in p2 favors the average-split test, while lack of the latter information in p6 favors the line-split test. If simplicity of a classifier is also considered, the decision tree learned with the average-split test from p2 would be judged as the best.

Secondly, we modified the selected period to 90, 180, 270, 360 days under an interpolation interval 5 days and the number of required medical tests per 30 days 1. We show the results in table 4. From the table, we see that the average-split test and the line-split test almost always outperform our standard-example split test in cost though there is no clear winner between them. We again attribute these to lack of information on shapes of time sequences and the number of examples. Our standard-example split test performs relatively well for 90 and 180 and this would be due to their relatively large numbers of examples. If simplicity of a classifier is also considered, the decision tree learned with the line-split test from 180 would be judged as the best.

Thirdly, we modified the interpolation intervals to 2, 4, $\cdots$, 10 days under a 180-day period and the required number of medical tests 6. We show the results in table 5 and 6. From the table 5, we see that our standard-example split test and the line-split test outperform the average-split test in cost though there is no clear winner between them. Since the 180 in table 4 represents 180BCp6i5, it would be displayed as i5 in this table. Our poor performance of cost 0.35 for i5 shows that our method exhibits good performance for small and large intervals, and this fact requires further investigation. If simplicity of a classifier is also considered, the line-split test is judged as the best and we again attribute this to lack of information for our method.

Lastly, we modified the Laplace correction parameter l to $0, 1, \cdots, 5$ under a 180-day period, the required number of medical tests 6, and a 6-day interpolation interval. We show the results in table 7. From the table, we see that the Laplace correction increases cost for our standard-example split test and the line-split test contrary to our expectation. Even for the average-split test, the case without the Laplace correction ($l = 0$) rivals the best case with the Laplace correction (l = 1). The table shows that these come from the fact that the Laplace correction lowers sensitivity but this requires further investigation.

3.3 Analysis of Experiments

In the experiments of [8], we employed longer time sequences and a larger number of training examples than in this paper. It should be also noted that the class ratio in [8] was nearly equivalent. We believe that our time-series decision tree is adequate for this kind of classification problems. The classification problems in this paper, since they neglect medical tests data after a biopsy, exhibit opposite characteristics, favoring a robust method such as the average-split test. Though it is appropriate to neglect medical tests data after a biopsy from medical viewpoint, the effect is negative for our time-series decision tree.

The decision trees which were constructed using the combined-split test and the average-split test contain many LC-leaves. Most of the leaves contain a small number of training examples, thus they rarely correspond to a test example. The high sensitivity and low cost exhibited by decision trees learned with the average-split test might come from their large sizes since they predict the LC class more frequently than a tree constructed with the combined-split test. This observation led us to consider modifying tree-structures in order to increase sensitivity and decrease cost.

4. Experiments for Goodness of a Split Test

4.1 Motivations

Table 8 Two examples of a split test

Split test	Left	Right	gain	gain ratio
test 1	6 (0, 6)	113 (76, 37)	0.077	0.268
test 2	47 (42, 5)	98 (34, 38)	0.122	0.160

From the discussions in the previous section, we considered to use the medical tests data before a biopsy and to replace gain ratio by gain. The former was realized by using the data sets employed in [8]. For the latter, consider their characteristics as goodness of a split test with tests 1 and 2 in table 8. Tests 1 and 2 are selected with gain ratio and gain respectively. As stated in [7], gain ratio tends to select an unbalanced split test where a child node has an extremely small number of examples. We believe that example 1 corresponds to this case and determined to perform a systematic comparison of the two criteria.

4.2 Experiments

We have compared our standard-example split test, the cluster-example split test, the average-split test, a method by Geurts [3], and a method by Kadous [5]. We settled N_{max} = 5 in the method of Geurts, and the number of discretized bins 5 and the number of clusters 5 in the method of Kadous. Experiments were performed with a leave-one-out method, and without the Laplace correction.

We show the results in table 9, and the decision trees learned from all data with the standard-example split test, the cluster-example split test, and the average-split test in figures 2, 3, and 4 respectively. The conditions are chosen so that each of them exhibits the lowest cost for the corresponding method.

From the table, we see that our standard-example split test performs better with gain ratio, and the cluster-example split test and the average-split test perform better with gain. We think that the former is due to affinity of gain ratio, which tends to select an unbalanced split, to our standard-example split test, which splits examples based on their similarities or dissimilarities to its standard example. Similarly, we think that the latter is due to affinity of gain, which is known to exhibit no such tendency, to the cluster-example split test and the average-split test, both of which consider characteristics of two children nodes in split. Actually, we see from the figures that a standard-example split test tends to produce a small-sized leaf while a cluster-example split test and an average-split test tend to construct a relatively balanced split.

Table 2 Data sets employed in the experiments

experiments	data (# of non-LC patients : # of LC patients)
experiments for the number of medical tests	180BCp6i5 (68:23), 180BCp3i5 (133:40), 180BCp2i5 (149:42)
experiments for the selected period	90BCp3i5 (120:38), 180BCp6i5 (68:23), 270BCp9i5 (39:15), 360BCp12i5 (18:13)
experiments for the interpolation interval	180BCp6i2, 180BCp6i4, 180BCp6i6, 180BCp6i8, 180BCp6i10 (all 68:23)

Table 3 Results of experiments for test numbers, where data sets p6, p3, and p2 represent 180BCp6i5, 180BCp3i5, and 180BCp2i5 respectively

	accuracy (%)			size			cost			sensitivity			specificity		
method	p6	p3	p2	p6	p3	p2	p6	p3	p2	p6	p3	p2	p6	p3	p2
Combined	78.0	75.7	80.6	10.9	20.5	18.9	0.35	0.35	0.33	0.52	0.53	0.52	0.87	0.83	0.89
Average	83.5	82.1	87.4	3.2	24.7	7.4	0.39	0.27	0.27	0.39	0.63	0.57	0.99	0.88	0.96
Line	84.6	82.7	85.9	9.0	22.7	3.6	0.30	0.27	0.34	0.57	0.63	0.43	0.94	0.89	0.98

Table 4 Results of experiments for periods, where data sets 90, 180, 270, and 360 represent 90BCp3i5, 180BCp6i5, 270BCp9i5, and 360BCp12i5 respectively

	accuracy (%)				size				cost				sensitivity				specificity			
method	90	180	270	360	90	180	270	360	90	180	270	360	90	180	270	360	90	180	270	360
Combined	77.8	78.0	64.8	45.2	19.5	10.9	8.5	5.5	0.36	0.35	0.52	0.69	0.50	0.52	0.33	0.23	0.87	0.87	0.77	0.61
Average	79.7	83.5	79.6	71.0	23.7	3.2	8.7	6.4	0.30	0.39	0.41	0.40	0.61	0.39	0.40	0.54	0.86	0.99	0.95	0.83
Line	77.2	84.6	74.1	48.4	18.7	9.0	8.7	6.5	0.41	0.30	0.40	0.58	0.39	0.57	0.47	0.38	0.89	0.94	0.85	0.56

Table 5 Results for accuracy, size, and cost of experiments for intervals, where data sets i2, i4, i6, i8, and i10 represent 180BCp6i2, 180BCp6i4, 180BCp6i6, 180BCp6i8, and 180BCp6i10 respectively

	accuracy (%)					size					cost				
method	i2	i4	i6	i8	i10	i2	i4	i6	i8	i10	i2	i4	i6	i8	i10
Combined	85.7	85.7	82.4	81.3	82.4	10.9	10.9	12.4	12.3	12.4	0.29	0.31	0.33	0.33	0.33
Average	84.6	84.6	83.5	84.6	82.4	3.0	3.0	3.2	3.9	5.1	0.36	0.36	0.39	0.36	0.39
Line	85.7	83.5	83.5	84.6	79.1	9.0	9.0	8.9	9.1	11.2	0.29	0.32	0.32	0.30	0.32

Table 6 Results for sensitivity and specificity of experiments for intervals

	sensitivity					specificity				
method	i2	i4	i6	i8	i10	i2	i4	i6	i8	i10
Combined	0.57	0.52	0.52	0.52	0.52	0.96	0.97	0.93	0.91	0.93
Average	0.43	0.43	0.39	0.43	0.39	0.99	0.99	0.99	0.99	0.97
Line	0.57	0.52	0.52	0.57	0.57	0.96	0.94	0.94	0.94	0.87

Table 7 Results of experiments for Laplace correction values with 180BCp6i6, where methods C, A, and L represent Combined, Average, and Line respectively

	accuracy (%)			size			cost			sensitivity			specificity		
value	C	A	L	C	A	L	C	A	L	C	A	L	C	A	L
0	86.8	85.7	82.4	10.9	10.8	7.4	0.28	0.29	0.33	0.57	0.57	0.52	0.97	0.96	0.93
1	82.4	83.5	83.5	12.4	3.2	8.9	0.33	0.39	0.32	0.52	0.39	0.52	0.93	0.99	0.94
2	81.3	83.5	80.2	9.1	3.0	9.0	0.36	0.39	0.38	0.48	0.39	0.43	0.93	0.99	0.93
3	83.5	73.6	83.5	9.1	2.5	9.0	0.30	0.63	0.34	0.57	0.00	0.48	0.93	0.99	0.96
4	81.3	83.5	79.1	9.2	2.6	8.9	0.36	0.39	0.39	0.48	0.39	0.43	0.93	0.99	0.91
5	82.4	83.5	82.4	9.1	2.7	8.9	0.35	0.39	0.37	0.48	0.39	0.43	0.94	0.99	0.96

Table 9 Experimental results with gain and gain ratio

		accuracy (%)		size		cost		sensitivity		specificity	
method	goodness	H1	H2	H1	H2	H1	H2	H1	H2	H1	H2
SE-split	gain	64.1	78.1	10.6	7.2	0.34	0.25	0.67	0.73	0.62	0.82
	gain ratio	79.7	85.9	9.0	7.1	0.24	0.18	0.73	0.80	0.85	0.91
CE-split	gain	81.2	76.6	9.0	8.7	0.20	0.23	0.80	0.77	0.82	0.76
	gain ratio	65.6	73.4	9.4	7.2	0.36	0.31	0.63	0.67	0.68	0.79
AV-split	gain	79.7	79.7	7.8	10.8	0.22	0.24	0.77	0.73	0.82	0.85
	gain ratio	73.4	70.3	10.9	11.4	0.31	0.39	0.67	0.57	0.79	0.82
Geurts	gain	68.8	70.3	10.1	9.7	0.28	0.32	0.73	0.67	0.65	0.74
	gain ratio	71.9	67.2	10.0	9.2	0.29	0.29	0.70	0.73	0.74	0.62
Kadous	gain	65.6	62.5	12.6	12.0	0.38	0.41	0.60	0.57	0.71	0.68
	gain ratio	71.9	65.6	8.8	13.2	0.29	0.27	0.70	0.77	0.74	0.56
1-NN		82.8	84.4	N/A	N/A	0.19	0.18	0.80	0.80	0.85	0.88

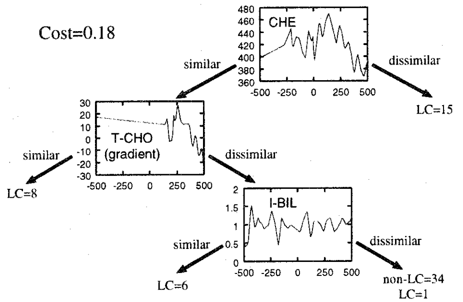

Figure 2 SE-split decision tree (H2, GainRatio)

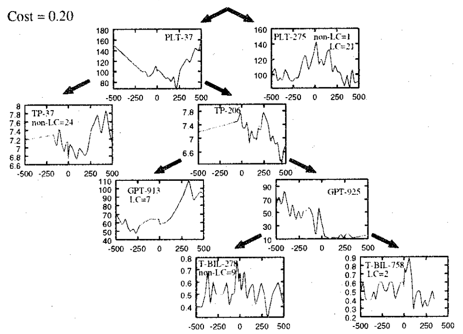

Figure 3 CE-split decision tree (H1, Gain)

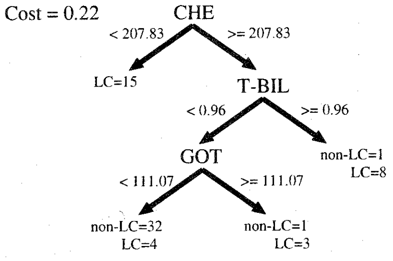

Figure 4 AV-split decision tree (H1, Gain)

5. Conclusions

For our time-series decision tree, we investigated the case in which medical tests before a biopsy are neglected and the case in which goodness of a split test is altered. In the former case, our time-series decision tree is outperformed by simpler decision trees in misclassification cost due to lack of information on sequences and examples. In the latter case, our standard-example split test performs better with gain ratio, and the cluster-example split test and the average-split test perform better with gain probably due to affinities in each combination. We plan to extend our approach as both a cost-sensitive learner and a discovery method.

References

[1] P. Berka: ECML/PKDD 2002 Discovery Challenge, Download Data about Hepatitis, *http://lisp.vse.cz/challenge/ ecmlpkdd2002/*, 2002 (current September 28th, 2002).

[2] J. P. Bradford, C. Kunz, R. Kohavi, C. Brunk, and C. E. Brodley: "Pruning Decision Trees with Misclassification Costs", *Proc. Tenth European Conference on Machine Learning (ECML)*, pp. 131–136, 1998.

[3] P. Geurts: "Pattern extraction for time series classification", *Principles of Data Mining and Knowledge Discovery (PKDD), LNAI 2168*, pp. 115–127, 2001.

[4] S. Hettich and S. D. Bay: The UCI KDD Archive *http://kdd.ics.uci.edu*, Irvine, CA: University of California, Department of Information and Computer Science, 1999.

[5] M. W. Kadous: "Learning comprehensible descriptions of multivariate time series", *Proc. Sixteenth International Conference on Machine Learning (ICML)*, pp. 454–463, 1999.

[6] E. J. Keogh: "Mining and Indexing Time Series Data", *Tutorial at the 2001 IEEE International Conference on Data Mining (ICDM)*, http://www.cs.ucr.edu/%7Eeamonn/ tutorial_on_time_series.ppt, 2001.

[7] J. R. Quinlan: *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, Calif., 1993.

[8] Y. Yamada, E. Suzuki, H. Yokoi, and K. Takabayashi: "Decision-tree Induction from Time-series Data Based on a Standard-example Split Test", *Proc. Twentieth International Conference on Machine Learning (ICML)*, 2003 (accepted for publication).