Minggang DONG, Ming LIU, Chao JING, 2022. One-against-all-based Hellinger distance decision tree for multiclass imbalanced learning. *Frontiers of Information Technology & Electronic Engineering*, 23(2):278-290.
https://doi.org/10.1631/FITEE.2000417

One-against-all-based Hellinger distance decision tree for multiclass imbalanced learning

Key words: Decision trees; Multiclass imbalanced learning; Node splitting criterion; Hellinger distance; One-against-all scheme

Corresponding author: Chao JING
E-mail: jingchao@glut.edu.cn
ORCID: https://orcid.org/0000-0002-4695-8746

Motivation

- We find that the Hellinger distance decision tree is inappropriate for solving the multiclass imbalanced classification problem. The main reason is that the Hellinger distance can solve only the two-class classification problem; it has limitations in identifying some differences in the multiclass imbalance problem.

- This process of computing the Hellinger distance does not tackle the issues of the multiclass distribution and the number of distinct classes, which are critical in the multiclass imbalance problem.

Contribution

- We introduce the scheme of one-against-all to the process of computing the splitting criterion of the one-against-all-based Hellinger distance (OAHD). During the computing process, we adopt the decomposition scheme so that OAHD can be extended to deal with the multiclass problem.

- We take into account the number of distinct classes and the distribution of the multiclass imbalance problem without considering the prior probability of the classes. Meanwhile, we modify the Gini index to incorporate it into the multiclass imbalance problem.

- We strictly prove that OAHD has the skew-insensitivity property, and that a purer node is sought by the splitting criterion.

The proposed splitting criterion

$$H_{\mathrm{D}}(j) = \sum_{v} \frac{\omega_{\theta}^{v}}{\omega_{\theta}^{u}} \exp\left(\sqrt{\frac{N_{\theta}^{v}}{N_{\theta}^{u}}} - \sqrt{\frac{N_{\bar{\theta}}^{v}}{N_{\bar{\theta}}^{u}}}\right)^{2} \cdot \frac{\beta^{u} - \beta^{v}}{\beta^{u}} \exp\left(\sum_{k=1}^{C} W_{k}^{2}\right),$$

where

$$W_k = \frac{\frac{N_k^v}{N_k^u}}{\sum_{i=1}^{C} \frac{N_i^v}{N_i^u}}, \quad \frac{\omega_{\theta}^{v}}{\omega_{\theta}^{u}} = \frac{\frac{N_{\theta}^{v}}{N_{\theta}^{u}} + \frac{N_{\bar{\theta}}^{v}}{N_{\bar{\theta}}^{u}}}{\sum_{v}\left(\frac{N_{\theta}^{v}}{N_{\theta}^{u}} + \frac{N_{\bar{\theta}}^{v}}{N_{\bar{\theta}}^{u}}\right)}.$$

Information of the imbalanced data sets

Table 1 Description of the imbalanced data sets, including the name, size, number of attributes, number of classes, class distribution, and imbalance ratio (IR)

Number	Name	Size	Number of attributes	Number of classes	Class distribution	IR
1	ESL12vs3vs456vs7vs89	488	4	5	62/14/351/38/23	25.1
2	Heart	270	13	2	150/120	1.3
3	Liver	345	6	2	145/200	1.4
4	Wine	178	13	3	71/59/48	1.5
5	Glass	214	10	6	76/70/29/17/13/9	8.5
6	Automobile12vs345vs6	205	71	3	25/153/27	6.1
7	ERA	1000	4	9	142/181/172/88/158/18/92/31/118	10.0
8	ERA1vs2345vs7vs8vs9	1000	4	5	771/88/92/31/18	42.8
9	Yeast52	982	8	5	463/25/35/429/30	18.5
10	Plates-faults1	1941	27	7	158/190/391/72/55/402/673	12.2
11	Plates-faults3	1941	27	5	158/863/793/72/55	15.7
12	abalone8discre	2148	10	8	126/203/267/487/634/259/115/57	11.1
13	abalone10discre	2297	10	10	57/115/259/391/634/487/126/103/67/58	11.1
14	page-blocks	559	10	4	329/115/87/28	11.8
15	pendigits	1100	16	10	115/114/114/106/114/106/105/115/105/106	1.1
16	housing5	506	13	5	36/123/239/77/31	7.7
17	vertebral-column	310	6	3	60/150/100	2.5
18	vehicle-mc	846	18	3	199/429/218	2.2
19	vowel5	990	10	5	180/90/360/270/90	4.0
20	vowel7	990	10	8	90/90/90/90/180/90/180/180	2.0

Experimental results of OAHD and comparisons

Table 2 Precision of the proposed OAHD and the five decision trees for the 20 imbalanced data sets

Number	Precision					
	CART	C4.5	DCSM	iHD	iHDw	OAHD
1	0.532 14	**0.721 43**	0.542 86	0.521 43	0.507 14	0.607 14
2	0.704 17	0.669 58	**0.719 17**	0.703 17	0.701 25	0.705 67
3	0.585 86	0.563 10	0.572 07	0.575 17	0.586 55	**0.612 41**
4	0.910 42	0.915 63	0.908 33	0.926 04	**0.927 08**	0.916 42
5	0.411 11	0.688 89	0.411 11	0.655 56	0.655 56	**0.716 67**
6	0.802 00	0.220 00	0.820 00	0.812 00	0.806 00	**0.840 00**
7	**0.777 78**	**0.777 78**	**0.777 78**	**0.777 78**	**0.777 78**	**0.777 78**
8	**0.777 78**	0.766 67	**0.777 78**	**0.777 78**	**0.777 78**	**0.777 78**
9	0.372 00	0.406 00	0.354 00	0.308 00	0.298 00	**0.452 00**
10	0.720 91	0.431 82	0.717 27	0.780 91	0.780 91	**0.814 55**
11	0.650 91	0.305 45	0.619 11	0.610 09	0.611 82	**0.706 36**
12	0.448 25	**0.516 67**	0.451 60	0.450 26	0.447 37	0.452 63
13	0.438 60	**0.556 14**	0.438 60	0.478 07	0.477 19	0.536 84
14	0.921 43	0.791 07	0.916 07	0.967 86	0.967 86	**0.978 57**
15	0.888 10	0.933 33	0.912 86	0.910 95	0.910 95	**0.935 71**
16	0.698 39	0.690 32	0.695 16	0.737 10	0.725 81	**0.803 23**
17	0.562 50	0.590 83	0.559 17	0.602 50	0.603 33	**0.632 50**
18	0.868 03	0.848 24	0.867 67	0.863 33	0.864 32	**0.887 19**
19	0.725 56	0.592 22	0.708 33	**0.796 11**	0.794 44	0.768 33
20	**0.933 33**	0.898 89	0.915 00	0.886 67	0.882 78	0.898 33

Best results are in bold

Experimental results of OAHD and comparisons

Table 3 F-measure of the proposed OAHD and the five decision trees for the 20 imbalanced data sets

Number	F-measure					
	CART	C4.5	DCSM	iHD	iHDw	OAHD
1	0.505 30	**0.655 61**	0.519 13	0.527 51	0.516 01	0.567 48
2	0.723 26	0.695 15	**0.731 30**	0.714 88	0.715 83	0.724 09
3	0.577 47	0.574 57	0.583 52	0.571 72	0.576 85	**0.592 08**
4	0.893 04	0.919 18	0.892 39	0.931 18	**0.931 23**	0.919 97
5	0.461 15	0.593 11	0.450 84	0.640 51	0.643 54	**0.712 75**
6	0.779 02	0.342 09	0.770 05	0.841 82	0.818 60	**0.881 63**
7	**0.700 00**	0.662 82	**0.700 00**	**0.700 00**	**0.700 00**	**0.700 00**
8	**0.700 00**	0.696 67	**0.700 00**	**0.700 00**	**0.700 00**	**0.700 00**
9	0.418 81	**0.528 59**	0.400 76	0.354 33	0.337 16	0.509 03
10	0.702 79	0.452 78	0.695 39	**0.755 38**	0.753 38	0.745 75
11	0.674 37	0.401 17	0.639 41	0.639 21	0.629 27	**0.718 38**
12	0.480 87	**0.515 25**	0.470 95	0.495 10	0.486 87	0.495 27
13	0.456 63	0.491 67	0.470 74	0.467 98	0.463 92	**0.496 88**
14	0.942 79	0.857 19	0.939 66	0.956 55	0.955 70	**0.974 97**
15	0.901 60	0.936 70	0.909 53	0.927 09	0.927 31	**0.938 30**
16	0.757 11	0.729 26	0.755 84	0.750 49	0.741 55	**0.792 73**
17	0.562 06	0.579 52	0.550 98	0.602 69	**0.607 41**	0.603 68
18	0.858 73	0.847 00	0.857 66	0.857 42	0.856 96	**0.865 68**
19	0.737 90	0.592 81	0.719 93	**0.799 36**	0.796 36	0.781 71
20	**0.933 34**	0.898 23	0.920 20	0.897 27	0.894 59	0.906 68

Best results are in bold

Experimental results of OAHD and comparisons

Table 4 MAUC of the proposed OAHD and the five decision trees for the 20 imbalanced data sets

Number	MAUC					
	CART	C4.5	DCSM	iHD	iHDw	OAHD
1	0.833 75	0.841 53	0.840 42	0.840 81	0.840 84	**0.843 43**
2	**0.802 71**	0.779 60	0.787 18	0.786 63	0.784 89	0.790 75
3	0.648 13	0.630 91	0.637 91	0.639 93	0.642 05	**0.648 26**
4	0.934 49	0.940 99	0.938 66	0.941 99	0.943 96	**0.944 73**
5	0.809 81	0.811 31	0.810 23	0.814 96	0.817 71	**0.818 41**
6	**0.905 52**	0.828 06	0.871 88	0.882 58	0.886 39	0.892 06
7	**0.710 23**	0.709 76	0.709 92	0.710 00	0.710 05	0.710 08
8	0.760 96	**0.761 34**	0.761 21	0.761 15	0.761 11	0.761 09
9	0.698 26	0.702 75	0.701 74	0.705 37	0.707 09	**0.709 42**
10	0.880 13	0.874 97	0.876 52	0.881 04	0.883 76	**0.884 13**
11	**0.852 54**	0.844 26	0.844 72	0.844 56	0.844 08	0.846 79
12	0.680 53	0.676 45	0.677 71	0.678 28	0.678 47	**0.681 56**
13	0.668 15	0.668 86	0.668 66	0.668 96	0.669 16	**0.671 26**
14	0.962 73	0.956 87	0.958 39	0.960 54	0.961 59	**0.963 35**
15	0.940 62	0.940 39	0.939 64	0.939 62	0.939 56	**0.945 98**
16	0.856 90	0.846 60	0.849 80	**0.857 64**	0.849 20	0.856 93
17	0.854 26	0.843 78	0.847 47	0.848 17	0.848 98	**0.860 09**
18	0.940 75	0.940 58	**0.946 08**	0.942 82	0.940 25	0.944 57
19	0.900 69	0.900 22	0.902 93	0.907 67	**0.915 32**	0.911 44
20	**0.920 14**	0.907 17	0.910 00	0.911 42	0.911 87	0.919 36

Best results are in bold

Mean ranking of OAHD and comparisons

Table 5 Mean ranking of the proposed OAHD and the five decision trees for the three metrics on the 20 imbalanced data sets

Algorithm	Mean ranking		
	Precision	F-measure	MAUC
CART	3.875	3.80	3.35
C4.5	4.175	4.40	4.90
DCSM	3.925	4.00	4.30
iHD	3.525	3.35	3.50
iHDw	3.725	3.60	3.35
OAHD	**1.775**	**1.85**	**1.60**

Best results are in bold

Results of Friedman and Nemenyi tests

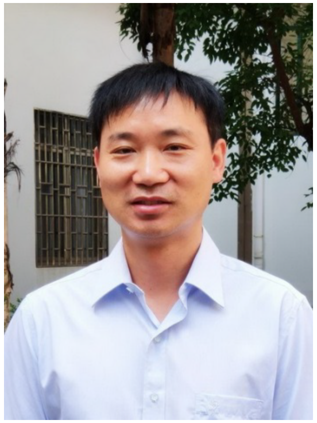
Table 6 Friedman test and Nemenyi test for the six methods on the 20 data sets, with OAHD as the base classifier

Algorithm	Precision	F-measure	MAUC
Friedman test	Reject	Reject	Reject
CART	✓	✓	✓
C4.5	✓	✓	✓
DCSM	✓	✓	✓
iHD	✓		✓
iHDw	✓	✓	✓
OAHD	Base	Base	Base

The symbol "✓" indicates that OAHD outperforms the compared algorithm

Conclusions

- In OAHD, considering the multiclass imbalance problem, the number of distinct classes and the class distribution are considered without a prior probability of the classes.
- We modify the Gini index to fit the multiclass imbalance problem, which ensures the purity of the node in the decision tree.
- We theoretically prove that the proposed splitting criterion enables the decision tree with the property of skew-insensitivity and the ability to seek a purer node.
- OAHD is compared with the five different unpruned decision trees upon 20 data sets. The experimental results show that the proposed splitting criterion is better than the five other splitting criteria. Moreover, the Friedman and Nemenyi tests are conducted to evaluate the performances of the six decision trees; the results demonstrate that this improvement is statistically significant.



董明刚，博士，教授。研究兴趣包括多目标优化、不平衡学习、目标检测等。2012年于浙江大学获博士学位，2016~2017年于美国罗德岛大学任访问学者。中国计算机学会高级会员。现就职于桂林理工大学信息科学与工程学院，主持国家自然科学基金2项、省部级项目2项，发表论文40余篇。

敬超，博士，副教授。研究兴趣包括人工智能与机器学习、大数据处理、云计算等。2014年于上海交通大学获博士学位。2017-2018年于美国布朗大学工程学院任访问学者。现就职于桂林理工大学信息科学与工程学院，主持/参与国家自然科学基金3项、省部级项目3项，发表论文20余篇，并担任国内外多种知名期刊、会议评审人。