scientific reports

Check for updates

OPEN

Explainable analytics: understanding causes, correcting errors, and achieving increasingly perfect accuracy from the nature of distinguishable patterns

Hao-Ting Pai[1,3]✉ & Chung-Chian Hsu[2,3]

In addition to pursuing accurate analytics, it is invaluable to clarify how and why inaccuracy exists. We propose a transparent classification (TC) method. In training, data consist of positive and negative observations. To obtain positive patterns, we find the intersection between each of the two positive observations. The negative patterns are obtained in the same manner. Next, pure positive and pure negative patterns are established by selecting patterns that appear in only one type. In testing, such pure positive and pure negative patterns are used for scoring observations. Next, an observation is classified as positive if its positive score is not zero or if both its positive and negative scores are zero; otherwise, it is classified as negative. By experiment, TC can identify all positive (e.g., malignant) observations at low ratios of training to testing data, e.g., 1:9 using the Breast Cancer Wisconsin (Original) and 3:7 using the Contraceptive Method Choice. Without fine-tuned parameters and random selection, the uncertainty of the methodology is eliminated when using TC. TC can visualize causes, and therefore, prediction errors in a network are traceable and can be corrected. Furthermore, TC shows potential in identifying whether the ground truth is incorrect (e.g., identifying diagnostic errors).

Accurate prediction plays a pivotal role in analytics; however, in reality, researchers usually face the challenge of explaining how and why a prediction is inaccurate[1,2]. According to a survey[3], outpatient diagnostic errors occur at a rate of 5.08% (approximately 12 million US adults) per year. Even a 1% reduction in errors can save the lives of millions of people. We consider three major types of errors. The first error type, *faults in data*, includes human mistakes or defective instrumentation, from which bad data is produced. Without domain knowledge, this type of fault is difficult to correct. Nevertheless, we should remove inconsistencies, i.e., when observations in the positive class are identical to those in the negative class. In addition, positive and negative observations may have similar patterns that are inextricably interwoven, e.g., people with similar profiles may exhibit different behaviors. Lim et al.[4] showed that the contraceptive method choice (CMC) dataset[5] is the most difficult to classify, and the minimum error rates are greater than 0.4. The second type of error is related to *mismatches between the data and the methods*. Data, which contain categorical (e.g., country), numerical (e.g., age), or both types of information, place natural constraints on the analysis. For categorical information, only the number of items and the mode are statistically relevant[6]. Therefore, a numerical-orientated method is inherently inadequate for categorical data. Numerical values can be transformed into categorical values by discretization[7], which is a technique that has been widely applied to knowledge discovery and data mining (KDD) applications[8]. However, bias occurs if categories are not representative of numerical values. The third error is the *big data challenge*, i.e., the complexity of data is determined by the number of rows and features (columns). Particularly, computation tasks increase rapidly with the number of features, which is known as the curse of dimensionality (CoD)[9]. To address CoD, dimension reduction and feature selection methods are utilized to reduce the complexity by extracting

[1]Bachelor Program of Big Data Applications in Business, National Pingtung University, Pingtung, Taiwan. [2]Department of Information Management, National Yunlin University of Science and Technology, Douliou, Taiwan. [3]International Graduate School of Artificial Intelligence, National Yunlin University of Science and Technology, Douliou, Taiwan. ✉email: htpai@mail.nptu.edu.tw

information that is practical for classification and cluster analysis. The extraction process, which is a trade-off between efficiency and effectiveness, may involve pruning large amounts of data. There may be pitfalls[10] in this process, and information related to errors may be missed.

Results

We conducted experiments with two public real-world datasets: the Breast Cancer Wisconsin (Original) (BCWO) and Contraceptive Method Choice (CMC) datasets, which are available in the UCI Machine Learning Repository[5]. Figure 1A,B show the results on the BCWO dataset, and Fig. 1C,D show the results on the CMC dataset. In Fig. 1A, perfect recall (i.e., a recall of 1.0) is achieved at the lowest ratio (i.e., 1:9) and 7 other ratios using the TC method. This means that this method is not only accurate for small amounts of data but is also stable when the amount of data increases. One error at the ratio of 2:8 and one error at the ratio of 3:7 occur because the positive observation PO_{223} is predicted as a negative observation. At the ratio of 4:6, PO_{223} is part of the training data but is not used in the testing data. Upon further exploration, at a ratio of 10:10[#], other than PO_{223}, the observations are irrelevant to the PPPs. Indeed, PO_{223} is related to the PPs, which is also relevant to the NOs. PPPs can eliminate the influence of PO_{223} on other observations. Novel observations are regarded as positive observations if they are unrelated to both PPPs and PNPs. Some NOs are predicted to be positive. As a result, a lower performance in terms of PR is observed when using the TC method. With an increased number of training observations, sufficient PPs or NPs may be obtained to provide accurate results and their causes, e.g., at the ratio of 1:9, PO_{627} has a novelty score of 70 by Rule 3; at the ratio of 2:8, PO_{627} contains 3 patterns in PPPs, namely, ['8|0.8'], ['6|1.5', '8|0.8'], and ['1|0.9', '8|0.8'], and has a positive score of 16 by Rule 1. Figure 1B shows that the suitable granularity of discretization provides a broad overview of how to discover more information. At a ratio of 1:9, one error, which is also caused by PO_{223}, results when using the TC method. At a ratio of 2:8, PO_{223} has 1 pattern ['3|0.0', '5|0.0'] in the PPPs and is related to PO_{33} and PO_{102}. At a ratio of 3:7, PO_{223} has 1 pattern ['3|0.0', '5|0.0'] in the PPPs and is related to PO_{33}, PO_{102}, and PO_{143}. At a ratio of 4:6, PO_{223} is a part of the training data and is not used in the testing data. As shown in Fig. 1C,D, at the ratios of 1:9 and 2:8, there are no negative observations in training (TRN); therefore, RE and PR are zero. At a ratio of 3:7, an insufficient amount of TRN leads to extreme results; a perfect RE but a relatively low PR are observed. The error rate is less than 0.4 when using a ratio of 7:3. In CMC, profile data are incapable of explaining behavior because (D) indicates that 33% (i.e., $(1473-980)/1473$) of cases are inherently inconsistent.

Discussion

In Fig. 2, we provide a visualization of the TC method. Compared to the provided images, the analysis of the categorical and numerical data have made it difficult to visualize how the causes are related to the results. Figure 2A shows that the TC method successfully predicts that O_{104} is positive because it is related to the six pure positive patterns that are obtained from the respective positive observations in training. Moreover, the thickness of lines represents the score and the degree of positiveness. Figure 2B shows the association between O_{104} and other observations. In the group containing O_{104}, the most common positive pattern is '0|1.3', and O_{24} and O_{57} have a significant influence on O_{104}. Figure 3A–C illustrates that the TC method is capable of addressing data with faulty class labels (ground truth) in terms of testing, training, and both testing and training. Figure 3D shows that the TC method can be utilized to correct errors in class labels. Specifically, it is observed that faulty class labels have extreme values of PS or NS, which becomes significant with an increasing amount of data. Based on the profound knowledge of most cases, the TC method could be useful for checking whether the original judgment (e.g., diagnosis) deserves further inspection.

In Fig. 4, a training data to testing data ratio of 5:5 is utilized so that the observations from input data #1 to #7 are used for training a model that is composed of pure positive patterns (PPPs) and pure negative patterns (PNPs). Next, the TC method utilizes the model to predict the class of a test observation. For example, for observation #8, the scores are obtained (i.e., NS = 4, PS = 0 and NT = 0); hence, the TC method predicts that #8 is negative. In particular, observation #12 is predicted to be positive by Rule 3, which shows that the TC method can be used to identify a novel case. In the area of machine learning, the training data to testing data ratio can show the performance of the proposed method. A method is considered excellent if it is accurate in a case that has a few training data but a large amount of testing data, e.g., a ratio of 2:8. Instead of a choosing a random selection, we select the observations based on their sequence so that the experimental results are reproducible. In addition, the ratios from 1:9 to 10:10[#] are implemented to provide a comprehensive view of the method. In this type of transparent manner, the TC method can help domain experts deeply understand the data.

According to Lim et al.[4], CMC has an inherent data quality problem, as the minimum error rate of the state-of-the-art methods is greater than 0.4. Although the minimum error rate of the TC method is 0.39, it has a limited ability to deal with this problem. In the TC method, the function of consistency can be used to identify observations that have identical patterns but different class labels to provide an interpretation of the minimum error rate. In social science data, we usually observe that people have identical profiles; however, their behaviors or decisions are quite different. Hou et al.[15] surveyed several analysis approaches of social media-based applications, which is useful for deeply exploring new significant factors in the classification task.

Methods

We propose a method of transparent classification, named TC, which not only strives to achieve accuracy but also clarifies the cause of inaccuracy. Furthermore, the design principles of the TC method to ensure reproducibility[11]. Figure 5 shows the processes of the TC method, and Fig. 4 provides a step-by-step approach to implementing the TC algorithms. In terms of *data preprocessing*, missing values and mixed values are addressed. Without randomness and reduction, information on the intrinsic nature of data is provided. In terms of *identifying distinguishable*

A

RA	RE	PR	AU	ER	TP	TN	FP	FN	TRP	TRN	TEP	TEN	UPP	PPP
1:9	1.000	0.701	0.905	0.140	206	335	88	0	35	35	206	423	251	201
2:8	0.994	0.777	0.976	0.093	178	329	51	1	62	78	179	380	592	468
3:7	0.993	0.760	0.988	0.100	152	288	48	1	88	122	153	336	1007	845
4:6	1.000	0.777	0.996	0.079	115	271	33	0	126	154	115	304	1630	1364
5:5	1.000	0.752	0.997	0.077	82	240	27	0	159	191	82	267	2266	1880
6:4	1.000	0.805	0.997	0.057	66	198	16	0	175	244	66	214	2599	2163
7:3	1.000	0.810	0.996	0.052	47	152	11	0	194	295	47	163	2968	2493
8:2	1.000	0.833	0.998	0.050	35	98	7	0	206	353	35	105	3225	2710
9:1	1.000	0.765	0.999	0.057	13	53	4	0	228	401	13	57	3723	3162
10:10#	1.000	1.000	1.000	0.000	241	458	0	0	241	458	241	458	4023	3422

RA: Ratios of training to testing; **RE**: Recall; **PR**: Precision; **AU**: AUC of ROC Curve; **ER**: Error rate; **TP**: True positive; **TN**: True negative; **FN**: False negative; **FP**: False positive; **TRP**: Positive observations of training; **TRN**: Negative observations of training; **TEP**: Positive observations of testing; **TEN**: Negative observations of testing; **UPP**: Unique positive patterns; **PPP**: Pure positive patterns. In RA, for comprehensive exploration, 10:10# means we respectively give the entire data in training and test.

B

RA	RE	PR	AU	ER	TP	TN	FP	FN	TRP	TRN	TEP	TEN	UPP	PPP
1:9	0.995	0.670	0.904	0.162	205	322	101	1	35	35	206	423	359	273
2:8	1.000	0.731	0.975	0.118	179	314	66	0	62	78	179	380	861	677
3:7	1.000	0.750	0.986	0.104	153	285	51	0	88	122	153	336	1390	1153
4:6	1.000	0.737	0.992	0.098	115	263	41	0	126	154	115	304	2308	1889
5:5	1.000	0.766	0.997	0.072	82	242	25	0	159	191	82	267	3211	2564
6:4	1.000	0.815	0.998	0.054	66	199	15	0	175	244	66	214	3666	2959
7:3	1.000	0.839	0.998	0.043	47	154	9	0	194	295	47	163	4162	3371
8:2	1.000	0.875	0.998	0.036	35	100	5	0	206	353	35	105	4555	3697
9:1	1.000	0.765	1.000	0.057	13	53	4	0	228	401	13	57	5286	4350
10:10#	1.000	1.000	1.000	0.000	241	458	0	0	241	458	241	458	5675	4687

C

RA	RE	PR	AU	ER	TP	TN	FP	FN	TRP	TRN	TEP	TEN	UPP	PPP
1:9	1.000	0.362	0.432	0.638	456	0	803	0	140	0	456	803	1539	1539
2:8	1.000	0.282	0.428	0.718	316	0	803	0	280	0	316	803	3555	3555
3:7	1.000	0.203	0.566	0.795	198	3	778	0	398	22	198	781	5539	4252
4:6	0.944	0.259	0.662	0.652	187	105	536	11	398	162	198	641	5539	2762
5:5	0.843	0.321	0.653	0.549	167	148	353	31	398	302	198	501	5539	2141
6:4	0.763	0.442	0.676	0.425	151	171	191	47	398	441	198	362	5539	1798
7:3	0.703	0.517	0.673	0.390	121	135	113	51	424	555	172	248	5938	1783
8:2	0.844	0.163	0.727	0.514	27	109	139	5	564	555	32	248	8345	2693
9:1	0.000	0.000	0.500	0.536	0	65	75	0	596	663	0	140	8744	2728
10:10#	1.000	1.000	1.000	0.000	596	803	0	0	596	803	596	803	8744	2566

D

RA	RE	PR	AU	ER	TP	TN	FP	FN	TRP	TRN	TEP	TEN	UPP	PPP
1:9	1.000	0.367	0.503	0.633	324	0	558	0	98	0	324	558	1480	1480
2:8	1.000	0.288	0.515	0.712	226	0	558	0	196	0	226	558	3524	3524
3:7	1.000	0.213	0.632	0.770	143	15	528	0	279	15	143	543	5580	4293
4:6	0.986	0.283	0.712	0.611	141	88	357	2	279	113	143	445	5580	2723
5:5	0.958	0.369	0.751	0.490	137	113	234	6	279	211	143	347	5580	1978
6:4	0.888	0.498	0.779	0.367	127	121	128	16	279	309	143	249	5580	1564
7:3	0.866	0.569	0.778	0.320	103	97	78	16	303	383	119	175	5994	1572
8:2	0.857	0.165	0.798	0.480	18	84	91	3	401	383	21	175	8192	2355
9:1	0.000	0.000	0.500	0.592	0	40	58	0	422	460	0	98	8560	2320
10:10#	1.000	1.000	1.000	0.000	422	558	0	0	422	558	422	558	8560	2120

Figure 1. Performance of TC in distinguishing between observations. (**A**, **B**) In Breast Cancer Wisconsin (Original) data set (BCWO), we map class values "malignant" to "1" and "benign" to "0". In case (**A**), the granularity of discretization is to the first decimal place, e.g., 1.68≈1.6, while in case (**B**) we take an integer for the granularity, e.g., 1.68≈1. (**C**, **D**) In Contraceptive Method Choice data set (CMC), we map class values "1 = No-use" to "1", "2 = Long-term" to "0", and "3 = Short-term" to "0". For (**C**) and (**D**), we set the same granularity as that of (**A**) and (**B**), respectively. For consistency, we remove observations that have identical features but different class labels. The number of observations is thus reduced from 1473 to 1399 in (**C**) and from 1473 to 980 in (**D**).

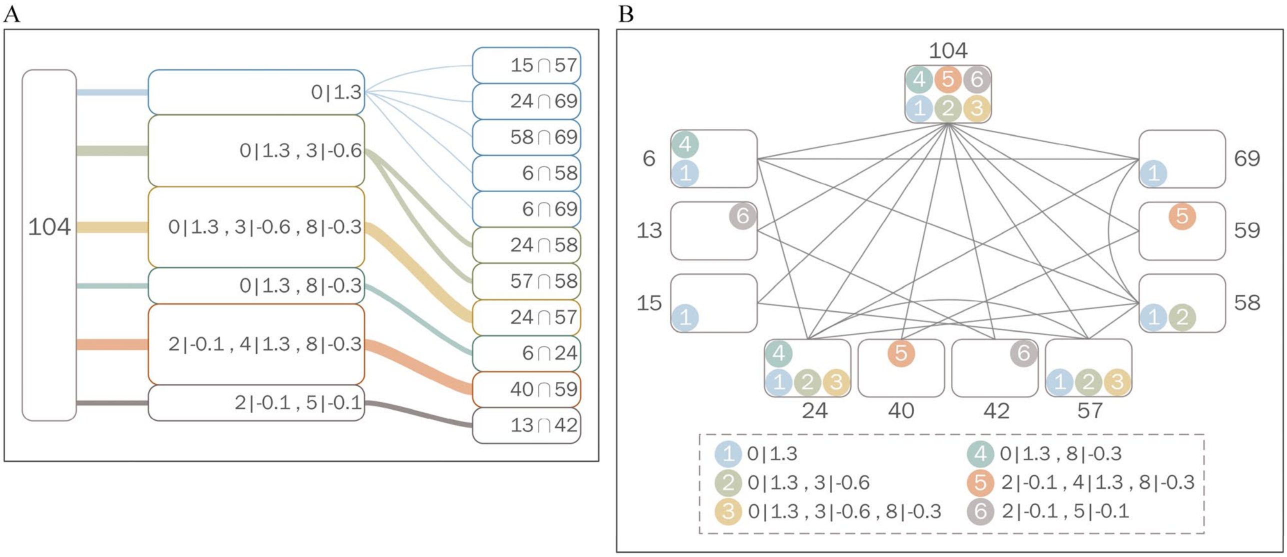

Figure 2. Association between patterns and observations in BCWO (**A**) at the ratio 1:9.

patterns, patterns are found from training observations, which are used for predicting which class a test observation belongs to, e.g., a malignant or benign tumor. By *increasing the ratios of training to testing observations*, the TC method represents a forest and the trees in the forest, and input data are given in sequence. To avoid CoD, patterns are found by intersecting pairwise observations in each of the classes, which possess essential features of miniature data. In the worst case, only $n \times (n-1)/2$ patterns are produced from n observations. In contrast, given the challenge of CoD, i.e., given the lowest threshold, k items yield 2^k item sets[12]; this challenge is encountered in KDD, and large amounts of item sets are pruned if the threshold is high. In terms of *positive patterns* (PPs), PPs are obtained from positive training observations (POs). For *pure PPs* (PPPs), any positive pattern that also appears in the negative training observations (NOs) are excluded. By set theory, an exclusion implies where none of the PPPs are included in any of the NOs, and hence the TC method can be used to distinguish between POs and NOs. Analogously, *negative patterns* (NPs) and *pure NPs* (PNPs) are the counterparts of PPs and PPPs. Without fine-tuned parameters or random selection, the uncertainty of the methodology is eliminated. In terms of *establishing the causes*, positive, negative, and novel degrees of a test observation O_t are accumulated by Rules 1, 2, and 3, respectively, which associate patterns with the observation and provide obvious clues for judgment.

$$\mathrm{PS_t} = \mathrm{PS_t} + \left|\mathrm{PO^{(ppp)}}\right| \times |\mathrm{ppp}|, \text{ if } \mathrm{ppp} \subseteq \mathrm{O_t} \tag{1}$$

$$\mathrm{NS_t} = \mathrm{NS_t} + \left|\mathrm{NO^{(pnp)}}\right| \times |\mathrm{pnp}|, \text{ if } \mathrm{pnp} \subseteq \mathrm{O_t} \tag{2}$$

$$\mathrm{NT_t} = |\mathrm{O_{TR}}|, \text{ if } \nexists_{i,j}, ppp_i \subseteq \mathrm{O_t}, pnp_j \subseteq \mathrm{O_t} \tag{3}$$

$$C(O_t) = \begin{cases} \text{Positive}, & \text{if } \mathrm{PS}_t > 0 \text{ or } \mathrm{NT}_t > 0 \\ \text{Negative}, & \text{otherwise} \end{cases} \tag{4}$$

In Rule 1, as shown in Formula (1), O_t containing patterns in the PPPs are given a positive score (PS). In Rule 2, as shown in Formula (2), O_t containing patterns in the PNPs are given a negative score (NS). In Rule 3, as shown Formula (3), O_t with no patterns in the PPPs or PNPs is considered novel, and the novelty score (NT) is equal to the number of training observations. The observation O_t is classified, as shown in Formula (4). Regarding notations, O_t is a testing observation, |ppp| is the cardinality of the pure positive pattern ppp, $|\mathrm{PO^{(ppp)}}|$ is the number of positive observations containing a ppp in the training set, $\mathrm{PS_t}$ is the positive score for O_t, $|\mathrm{O_{TR}}|$ indicates the number of observations in the training set, and $\mathrm{NT_t}$ is the novelty score for O_t.

To *understand the results of the analytics*, we evaluate the performance of the TC method by three measures: *precision*, *recall*, and the *area under curve (AUC)*[8,12]. According to the standards of diagnostic medicine[13]: AUC = 0.5, no discrimination; 0.7 ≤ AUC < 0.8, acceptable; 0.8 ≤ AUC < 0.9, excellent; and 0 9 ≤ AUC ≤ 1, outstanding. When evaluating the *causes for prediction errors*, error 1 (false-positive[14]), which is denoted as is denoted as NO_{t^*}, occurs if O_t is predicted as positive but is actually negative. In cause 1.1, although NO_{t^*} contains pure positive patterns, it should not. In cause 1.2, NO_{t^*} is novel, namely, it has no patterns in the PPPs and PNPs. Error 2 (false negative[14]), which is denoted as PO_{t^*}, occurs if O_t is predicted to be negative but is actually positive. In cause 2.1, PO_{t^*} contains pure negative patterns but has no pure positive patterns, although it should. Prediction errors occur due to insufficient training data or labeling errors in the training data. Increasing the number of

A

RA	RE	PR	AU	ER	TP	TN	FP	FN	TRP	TRN	TEP	TEN	UPP	PPP
1:9	0.995	0.660	0.904	0.167	202	322	104	1	35	35	203	426	359	273
2:8	1.000	0.718	0.975	0.123	176	314	69	0	62	78	176	383	861	677
3:7	1.000	0.735	0.985	0.110	150	285	54	0	88	122	150	339	1390	1153
4:6	1.000	0.718	0.991	0.105	112	263	44	0	126	154	112	307	2308	1889
5:5	1.000	0.738	0.996	0.080	79	242	28	0	159	191	79	270	3211	2564
6:4	1.000	0.778	0.996	0.064	63	199	18	0	175	244	63	217	3666	2959
7:3	1.000	0.786	0.996	0.057	44	154	12	0	194	295	44	166	4162	3371
8:2	1.000	0.800	0.995	0.057	32	100	8	0	206	353	32	108	4555	3697
9:1	1.000	0.588	0.985	0.100	10	53	7	0	228	401	10	60	5286	4350
10:10#	1.000	1.000	1.000	0.000	238	461	0	0	238	461	238	461	5600	4525

B

RA	RE	PR	AU	ER	TP	TN	FP	FN	TRP	TRN	TEP	TEN	UPP	PPP
1:9	0.981	0.699	0.908	0.145	202	336	87	4	32	38	206	423	318	219
2:8	1.000	0.762	0.980	0.100	179	324	56	0	59	81	179	380	806	590
3:7	1.000	0.789	0.989	0.084	153	295	41	0	85	125	153	336	1318	1040
4:6	1.000	0.777	0.994	0.079	115	271	33	0	123	157	115	304	2226	1760
5:5	1.000	0.781	0.997	0.066	82	244	23	0	156	194	82	267	3120	2416
6:4	1.000	0.835	0.998	0.046	66	201	13	0	172	247	66	214	3567	2798
7:3	1.000	0.839	0.997	0.043	47	154	9	0	191	298	47	163	4060	3202
8:2	1.000	0.875	0.998	0.036	35	100	5	0	203	356	35	105	4451	3522
9:1	1.000	0.765	1.000	0.057	13	53	4	0	225	404	13	57	5190	4153
10:10#	1.000	1.000	1.000	0.000	238	461	0	0	238	461	238	461	5578	4485

C

RA	RE	PR	AU	ER	TP	TN	FP	FN	TRP	TRN	TEP	TEN	UPP	PPP
1:9	0.980	0.689	0.907	0.149	199	336	90	4	32	38	203	426	318	219
2:8	1.000	0.749	0.980	0.106	176	324	59	0	59	81	176	383	806	590
3:7	1.000	0.773	0.988	0.090	150	295	44	0	85	125	150	339	1318	1040
4:6	1.000	0.757	0.992	0.086	112	271	36	0	123	157	112	307	2226	1760
5:5	1.000	0.752	0.996	0.074	79	244	26	0	156	194	79	270	3120	2416
6:4	1.000	0.797	0.996	0.057	63	201	16	0	172	247	63	217	3567	2798
7:3	1.000	0.786	0.994	0.057	44	154	12	0	191	298	44	166	4060	3202
8:2	1.000	0.800	0.993	0.057	32	100	8	0	203	356	32	108	4451	3522
9:1	1.000	0.588	0.985	0.100	10	53	7	0	225	404	10	60	5190	4153
10:10#	1.000	1.000	1.000	0.000	235	464	0	0	235	464	235	464	5504	4331

D

Faults in testing (cases 697, 698, and 699)					Faults in training (cases 6, 13, and 15)					Faults in training and testing (cases 6, 13, 15, 697, 698, and 699)				
RA	O	NT	PS	NS	RA	O	NT	PS	NS	RA	O	NT	PS	NS
1:9	697	0	42	0	1:9	223	0	0	50	1:9	697	0	33	0
1:9	698	0	9	0	1:9	286	0	0	13	1:9	698	0	9	0
1:9	699	0	23	0	1:9	345	0	0	4	1:9	699	0	10	0
...	...	...	...	...	1:9	569	0	0	4	...	...	...	...	...
9:1	697	0	2672	0	2:8	223	0	4	875	9:1	697	0	2293	0
9:1	698	0	134	0	2:8	286	0	123	9	9:1	698	0	134	0
9:1	699	0	1016	9	2:8	345	0	12	0	9:1	699	0	1016	9
10:10#	695	0	0	95204	2:8	569	0	13	0	10:10#	695	0	0	95204
10:10#	696	0	0	3537383	10:10#	5	0	0	143716	10:10#	696	0	0	3537383
10:10#	697	0	0	99	10:10#	6	0	0	81	10:10#	697	0	0	117
10:10#	698	0	0	106	10:10#	13	0	0	726	10:10#	698	0	0	106
10:10#	699	0	0	140	10:10#	15	0	0	81	10:10#	699	0	0	140

Figure 3. Tolerance to faulty class labels in BCWO (**B**). (**A**) For the case of testing, we change class labels of observations (i.e., 697, 698, and 699) from "1" to "0". (**B**) For the case of training, class labels of observations (i.e., 6, 13, and 15) are changed from "1" to "0". (**C**) For both cases, we change class labels of observations (i.e., 6, 13, 15, 697, 698, and 699) from "1" to "0". (**D**) Although faults in testing, TC can still classify O_{697}, O_{698}, and O_{699} as PO since the ratio 1:9. Regarding faults in training, the change results in an additional three errors at the ratio 1:9, i.e., O_{286}, O_{345}, and O_{569}. Since the ratio 2:8, the three errors are eliminated due to increased training data. Although faults in testing and training, TC can still classify O_{697}, O_{698}, and O_{699} as PO since the ratio 1:9. Note at the ratio 10:10#, O_{697}, O_{698}, and O_{699} belong to training data so that their PS are zero; nevertheless, we can identify them by NS.

A

Raw Data

Instance				Class
C0	C1	C2	C3	
a1	b1	1.1	0.1	0
a1	b1	1.0	0.1	1
a1	b1	0.9	1.9	0
a1	b1		0.2	0
a1	b2	1.0	0.1	0
a1	b2	9.7	1.8	1
a2	b1	9.9	1.9	1
a2	b2	9.8	1.8	1
a2	b1	1.1	0.1	0
a1	b2	9.6	0.2	0
a1	b2	1.0	1.9	0
a1	b1		1.8	0
a2	b2	1.2	0.2	0
a2	b1	9.9	0.1	0
a2	b1	8.9	1.7	0
a2	b2	0.9	1.8	1
a2	b2	9.5	1.9	1
a2	b2	9.6	1.8	1

Data Pre-processing

Exception	Discretization		Formation				Consistency					Data Input				
C2	C2	C3	C0	C1	C2	C3	C0	C1	C2	C3	Class	C0	C1	C2	C3	Class
1.1	(1.1-5.3)/4.3=-1	-1	0|a1	1|b1	2|-1	3|-1	~~0~~	~~2~~	~~4~~	~~7~~	~~0~~	0	2	4	8	0
1.0	-1	-1	0|a1	1|b1	2|-1	3|-1	~~0~~	~~2~~	~~4~~	~~7~~	~~1~~	0	2	6	7	0
0.9	-1	1	0|a1	1|b1	2|-1	3|1	0	2	4	8	0	0	3	4	7	0
NA	NA	-1	0|a1	1|b1	2|NA	3|-1	0	2	6	7	0	0	3	5	8	1
1.0	-1	-1	0|a1	1|b2	2|-1	3|-1	0	3	4	7	0	1	3	5	8	1
9.7	1	1	0|a1	1|b2	2|1	3|1	0	3	5	8	1	1	2	4	7	0
9.9	1	1	0|a2	1|b1	2|1	3|1	~~1~~	~~2~~	~~5~~	~~8~~	~~1~~	0	3	5	7	0
9.8	1	1	0|a2	1|b2	2|1	3|1	1	3	5	8	1	0	3	4	8	0
1.1	-1	-1	0|a2	1|b1	2|-1	3|-1	1	2	4	7	0	0	2	6	8	0
9.6	1	-1	0|a1	1|b2	2|1	3|-1	0	3	5	7	0	1	3	4	7	0
1.0	-1	1	0|a1	1|b2	2|-1	3|1	0	3	4	8	0	1	2	5	7	0
NA	NA	1	0|a1	1|b1	2|NA	3|1	0	2	6	8	0	1	3	4	8	1
1.2	-1	-1	0|a2	1|b2	2|-1	3|-1	1	3	4	7	0	1	3	5	8	1
9.9	1	-1	0|a2	1|b1	2|1	3|-1	1	2	5	7	0	1	3	5	8	1
8.9	1	1	0|a2	1|b1	2|1	3|1	~~1~~	~~2~~	~~5~~	~~8~~	~~0~~					
0.9	-1	1	0|a2	1|b2	2|-1	3|1	1	3	4	8	1					
9.5	1	1	0|a2	1|b2	2|1	3|1	1	3	5	8	1					
9.6	1	1	0|a2	1|b2	2|1	3|1	1	3	5	8	1					

B

Identifying Distinguishable Patterns

Ratio of training to testing 5 : 5

Training				
C0	C1	C2	C3	Class
0	2	4	8	0
0	2	6	7	0
0	3	4	7	0
0	3	5	8	1
1	3	5	8	1
1	2	4	7	0
0	3	5	7	0

Testing			
0	3	4	8
0	2	6	8
1	3	4	7
1	2	5	7
1	3	4	8
1	3	5	8
1	3	5	8

Negative Patterns (NP)	Pure NP (PNP)
{0,2,4,8},{0,2,6,7},{0,3,4,7},{1,2,4,7},{0,3,5,7},	{0,2,4,8},{0,2,6,7},{0,3,4,7},{1,2,4,7},{0,3,5,7},
{0,2,4,8}∩{0,2,6,7}={0,2}, {0,2,4,8}∩{0,3,4,7}={0,4}, {0,2,4,8}∩{1,2,4,7}={2,4}, {0,2,4,8}∩{0,3,5,7}={0}, {0,2,6,7}∩{0,3,4,7}={0,7}, {0,2,6,7}∩{1,2,4,7}={2,7}, {0,2,6,7}∩{0,3,5,7}={0,7}, {0,3,4,7}∩{1,2,4,7}={4,7}, {0,3,4,7}∩{0,3,5,7}={0,3,7}, {1,2,4,7}∩{0,3,5,7}={7}.	{0,2}∉{0,3,5,8},{0,2}∉{1,3,5,8}, {0,4}∉{0,3,5,8},{0,4}∉{1,3,5,8}, {2,4}∉{0,3,5,8},{2,4}∉{1,3,5,8}, ~~{0}∈{0,3,5,8}~~, {0,7}∉{0,3,5,8},{0,7}∉{1,3,5,8}, {2,7}∉{0,3,5,8},{2,7}∉{1,3,5,8}, {4,7}∉{0,3,5,8},{4,7}∉{1,3,5,8}, {0,3,7}∉{0,3,5,8},{0,3,7}∉{1,3,5,8}, {7}∉{0,3,5,8},{7}∉{1,3,5,8}.

Positive Patterns (PP)	Pure PP (PPP)
{0,3,5,8},{1,3,5,8},	{0,3,5,8},{1,3,5,8},
{0,3,5,8}∩{1,3,5,8}={3,5,8}.	{3,5,8}∉{0,2,4,8},{3,5,8}∉{0,2,6,7}, {3,5,8}∉{0,3,4,7},{3,5,8}∉{1,2,4,7}, {3,5,8}∉{0,3,5,7}.

C

Establishing the Causes

PNP	
Pattern	Score
{0,2,4,8}	4×4=16
{0,2,6,7}	4×4=16
{0,3,4,7}	4×4=16
{1,2,4,7}	4×4=16
{0,3,5,7}	4×4=16
{0,2}	2×2=4
{0,4}	2×2=4
{2,4}	2×2=4
{0,7}	2×2=4
{2,7}	2×2=4
{4,7}	2×2=4
{0,3,7}	3×3=9
{7}	1×1=1

Testing			
NS	Instance	NT	PS
4	{0,3,4,8}	0	0
4	{0,2,6,8}	0	0
5	{1,3,4,7}	0	0
5	{1,2,5,7}	0	0
0	{1,3,4,8}	7*	0
0	{1,3,5,8}	0	25
0	{1,3,5,8}	0	25

PPP	
Pattern	Score
{0,3,5,8}	4×4=16
{1,3,5,8}	4×4=16
{3,5,8}	3×3=9

*Note that: By the rule 3, the instance {1, 3, 4, 8} gets NS=0 and PS=0, and thus its NT's score is the number of instances of training.

D

Understanding Results of Analytics

Testing		Class (Ground Truth)
Instance	Prediction by TC	
{0,3,4,8}	0	0
{0,2,6,8}	0	0
{1,3,4,7}	0	0
{1,2,5,7}	0	0
{1,3,4,8}	1	1
{1,3,5,8}	1	1
{1,3,5,8}	1	1

Performance evaluation	
Precision	1.0
Recall	1.0
AUC of ROC Curve	1.0

Figure 4. Illustration of TC: a step-by-step approach.

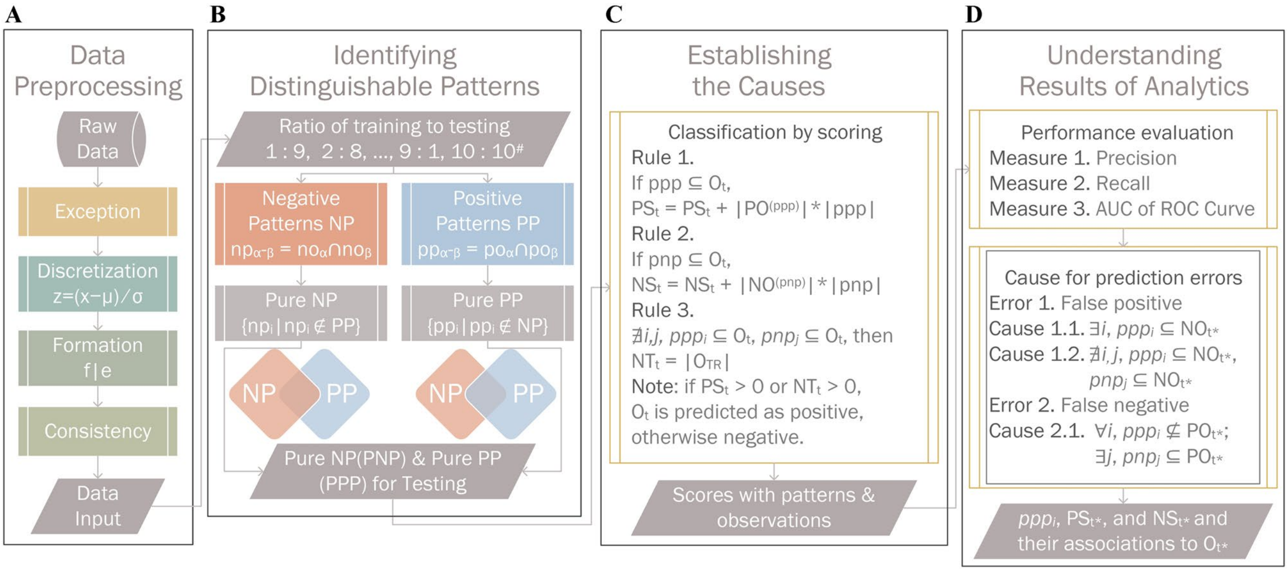

Figure 5. Processes of the transparent classification. (**A**) *Exception* treats missing values as categorical values instead of guesses. *Discretization* transfers numerical values to categorical ones by using z-score, where x are numerical values of a feature, μ is the mean and σ is the standard deviation. *Formation* defines the relations between features f and categorical values e. *Consistency* removes the contradictory observations that have identical features but different class labels. (**B**) *Ratio of training to testing* divides data into two parts: training and testing. In training, observations are split positive observations PO and negative observations NO. A *positive pattern* $pp_{\alpha\text{-}\beta}$ was discovered by intersecting po_α and po_β, where $\alpha = 1, 2, \ldots, n$ and $\beta = \alpha, \alpha+1, \ldots, n$, e.g., given $n = 3$, then $pp_{1\text{-}1} = po_1 \cap po_1 = \{f_1|e_{1,1}, f_2|e_{2,2}, f_3|e_{3,1}\}$, $pp_{1\text{-}2} = po_1 \cap po_2 = \{f_1|e_{1,1}, f_2|e_{2,2}, f_3|e_{3,1}\} \cap \{f_1|e_{1,1}, f_2|e_{2,1}, f_3|e_{3,1}\} = \{f_1|e_{1,1}, f_3|e_{3,1}\}$, $pp_{1\text{-}3} = po_1 \cap po_3$, $pp_{2\text{-}2} = po_2$, $pp_{2\text{-}3} = po_2 \cap po_3$, and $pp_{3\text{-}3} = po_3$. Note positive observations themselves are positive patterns, e.g., $pp_{1\text{-}1}$. A *negative pattern* $np_{\alpha\text{-}\beta}$ was found by $no_\alpha \cap no_\beta$, and negative observations themselves are negative patterns. *Pure PP* (*PPP*) excludes $pp_{\alpha\text{-}\beta}$ that appears in any negative observation. *Pure NP* (*PNP*) excludes $np_{\alpha\text{-}\beta}$ that appears in any positive observation. (**C**) In testing of an observation O_t, classification by scoring produces five outputs: *PS*, *NS*, *NT*, *PSP*, and *PSO*. *PS* stores observation's positive score by Rule 1: O_t contains a pure positive pattern *ppp*, increase *PS* by the number of features in *ppp*, multiplied by the number of positive observations containing *ppp*. Rule 2: if O_t contains a pure negative pattern *pnp*, increase *NS* by the number of features in *pnp* multiplied by the number of negative observations containing *pnp*. Rule 3: if O_t does not contain any *ppp* and *pnp*, assign *NT* to the number of training observations. *PSP* stores $ppp_{\alpha\text{-}\beta}$ related to O_t. *PSO* stores the training observations which contain $ppp_{\alpha\text{-}\beta}$. If *PS* or *NT* is greater than 0, classify O_t as positive otherwise negative. (**D**) *Performance evaluation* demonstrates the accuracy of TC. *Cause for prediction errors*, based on set theory, provides rational explanations for errors caused by TC.

training data helps to reduce prediction errors. If the portion of labeling errors is small, the TC method has the potential to identify labeling errors. Specifically, false negatives usually have a small NS.

Received: 5 December 2021; Accepted: 31 August 2022
Published online: 01 November 2022

References

1. Description, prediction, explanation. *Nat. Hum. Behav.* **5**, 1261 (2021). https://doi.org/10.1038/s41562-021-01230-5.
2. Gunning, D. *et al.* XAI—Explainable artificial intelligence. *Sci. Robot.* **4**, eaay7120 (2019).
3. Singh, H., Meyer, A. N. & Thomas, E. J. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. *BMJ Qual. Saf.* **23**, 727–731 (2014).
4. Lim, T. S., Loh, W. Y. & Shih, Y. S. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Mach. Learn.* **40**, 203–228 (2000).
5. D. Dheeru, C. Graff, UCI machine learning repository. http://archive.ics.uci.edu/ml (2019).
6. Stevens, S. S. On the theory of scales of measurement. *Science* **103**, 677–680 (1946).
7. Garcia, S., Luengo, J., Sáez, J. A., Lopez, V. & Herrera, F. A survey of discretization techniques: taxonomy and empirical analysis in supervised learning. *IEEE Trans. Knowl. Data Eng.* **25**, 734–750 (2012).
8. Tan, P. N., Steinbach, M. & Kumar, V. *Introduction to Data Mining* (Pearson, 2020).
9. Altman, N. & Krzywinski, M. The curse(s) of dimensionality. *Nat. Methods.* **15**, 399–400 (2018).
10. Jain, A. & Zongker, D. Feature selection: evaluation, application, and small sample performance. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**, 153–158 (1997).
11. McNutt, M. Reproducibility. *Science* **343**, 229 (2014).
12. Zaki, M. J. & Meira, W. Jr. *Data Mining and Machine Learning: Fundamental Concepts and Algorithms* (Cambridge University Press, 2020).
13. Mandrekar, J. N. Receiver operating characteristic curve in diagnostic test assessment. *J. Thorac. Oncol.* **5**, 1315–1316 (2010).
14. National Cancer Institute. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/false-positive-test-result (2021).

15. Hou, Q., Han, M. & Cai, Z. Survey on data analysis in social media: a practical application aspect. *Big Data Min. Anal.* **3**(4), 259–279 (2020).

Acknowledgements
We thank the editor and reviewers for offering comments on the manuscript.

Author contributions
Conceptualization: H.T.P. conceived the idea and C.C.H. provided for guidance. Methodology: H.T.P. Experiments: H.T.P. Visualization: H.T.P. Funding acquisition: H.T.P., C.C.H. Writing—original draft: H.T.P. Writing—review and editing: H.T.P., C.C.H.

Funding
This work was supported by the Ministry of Science and Technology (MOST) in Taiwan. MOST (Grant No. 110-2410-H-224-037-): H.T.P. MOST (Grant No. 110-2410-H-224-009): C.C.H.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to H.-T.P.