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

Hao‑Ting Pai1,3 & Chung‑Chian Hsu2,3

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 inaccurate1,2. According to a survey3, 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) dataset5 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 relevant6. Therefore, a numerical‑orientated method is inherently inadequate for categorical data. Numerical values can be transformed into categorical values by discretization7, which is a technique that has been widely applied to knowledge discovery and data mining (KDD) applications8. 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...
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\(^9\) 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\(^2\). Figure 1A,B show the results on the BCWO dataset, and Fig. 1C,D show the results on the CMC dataset. In Fig. 1A, a 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\(104\) is predicted as a negative observation. At the ratio of 4:6, PO\(104\) is part of the training data but is not used in the testing data. Upon further exploration, at a ratio of 10:10\(^4\), other than PO\(223\), the observations are irrelevant to the PPs. Indeed, PO\(223\) is related to the PPVs, 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 PPVs or NPVs may be obtained to provide accurate results and their causes, i.e., at the ratio of 1:9, PO\(104\) has a novelty score of 70 by Rule 3 at the ratio of 2:8, PO\(104\) 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\(143\) 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\(143\), PO\(102\), and PO\(104\). 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 and a 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\(223\) 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 NS=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 novel cases. 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 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\(^4\) 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.\(^5\) 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\(^4\). 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
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.

**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).
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; 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.

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).

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

$$NS_t = NS_t + \left| NO^{PNP}\right| \times \left| pnp\right|, \text{ if } pnp \subseteq O_t$$  \hspace{1cm} (2)

$$NT_t = |O_{TR}|, \text{ if } \forall i, j, ppp_i \subseteq O_t, pnp_j \subseteq O_t$$  \hspace{1cm} (3)

$$C(O_t) = \begin{cases} \text{Positive, if } PS_t > 0 \text{ or } NT_t > 0 \\ \text{Negative, otherwise} \end{cases}$$  \hspace{1cm} (4)

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). According to the standards of diagnostic medicine, 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), which is denoted as is denoted as $NO_+$, occurs if $O_t$ is predicted as positive but is actually negative. In cause 1.1, although $NO_+$ contains pure positive patterns, it should not. In cause 1.2, $NO_-$ is novel, namely, it has no patterns in the PPPs and PNPs. Error 2 (false negative, which is denoted as $PO_-$, occurs if $O_t$ is predicted to be negative but is actually positive. In cause 2.1, $PO_-$ 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
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 10:10#, 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.
Figure 4. Illustration of TC: a step-by-step approach.
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.

 recibed: 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/s41598-022-19650-2
2. Gunning, D. et al. 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. Ghara, C. Graf. 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 evaluation 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, 2005).
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. Zak, 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.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022