BMC Medical Informatics and Decision Making

RESEARCH ARTICLE

Open Access

Explanation and prediction of clinical data with imbalanced class distribution based on pattern discovery and disentanglement

Pei-Yuan Zhou* and Andrew K. C. Wong

Abstract

Background: Statistical data analysis, especially the advanced machine learning (ML) methods, have attracted considerable interest in clinical practices. We are looking for interpretability of the diagnostic/prognostic results that will bring confidence to doctors, patients and their relatives in therapeutics and clinical practice. When datasets are imbalanced in diagnostic categories, we notice that the ordinary ML methods might produce results overwhelmed by the majority classes diminishing prediction accuracy. Hence, it needs methods that could produce explicit transparent and interpretable results in decision-making, without sacrificing accuracy, even for data with imbalanced groups.

Methods: In order to interpret the clinical patterns and conduct diagnostic prediction of patients with high accuracy, we develop a novel method, Pattern Discovery and Disentanglement for Clinical Data Analysis (cPDD), which is able to discover patterns (correlated traits/indicants) and use them to classify clinical data even if the class distribution is imbalanced. In the most general setting, a relational dataset is a large table such that each column represents an attribute (trait/indicant), and each row contains a set of attribute values (AVs) of an entity (patient). Compared to the existing pattern discovery approaches, cPDD can discover a small succinct set of statistically significant high-order patterns from clinical data for interpreting and predicting the disease class of the patients even with groups small and rare.

Results: Experiments on synthetic and thoracic clinical dataset showed that cPDD can 1) discover a smaller set of succinct significant patterns compared to other existing pattern discovery methods; 2) allow the users to interpret succinct sets of patterns coming from uncorrelated sources, even the groups are rare/small; and 3) obtain better performance in prediction compared to other interpretable classification approaches.

Conclusions: In conclusion, cPDD discovers fewer patterns with greater comprehensive coverage to improve the interpretability of patterns discovered. Experimental results on synthetic data validated that cPDD discovers all patterns implanted in the data, displays them precisely and succinctly with statistical support for interpretation and prediction, a capability which the traditional ML methods lack. The success of cPDD as a novel interpretable method in solving the imbalanced class problem shows its great potential to clinical data analysis for years to come.

Keywords: Pattern discovery, Disentanglement, Clinical decision-making, Imbalance classification

Background

Clinical diagnostic decisions have a direct impact on the outcomes and treatment of patients in the clinical setting. As large volumes of biomedical and clinical data are being collected and becoming available for analysis, there is an increasing interest and need in applying

*Correspondence: choupeiyuan.ca@gmail.com
Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada

machine learning (ML) methods to diagnose diseases, predict patient outcomes and propose therapeutic treatments. Today, Deep Learning (DL) has been successful in assisting analysis and classifying medical scans, especially those forms of visual data. However, when dealing with relational datasets where no explicit pattern (except the class label if given) could be extracted from the input data to relate to the decision targets, the ML/DL process remains opaque. In addition, existing ensemble algorithms, such as Boosted SVM, or Random Forest could produce good predictive results, but the underlying patterns in support of the decision are still opaque and uninterpretable for the clinicians [1]. Hence, existing ML approaches on relational data are still encountering difficult problems concerning transparency, low data volume, and imbalance classes [2, 3].

To render transparency and interpretability, Decision Tree, Frequent Pattern Mining or Pattern Discovery were proposed. For decades, *Frequent Pattern Mining* [4–6] is an essential data mining task to discover knowledge in the form of association rules from relational data [6]. The association rules or patterns are made up of co-occurring items or attribute values (AVs) referred to as Attribute Value Associations (AVAs). However, as revealed in our recent work [7–9], the AVA forming patterns of different classes/targets could be overlapping or entangling with each other due to multiple entwining functional characteristics or factors of different groups/classes inherent in the source environments. For example, in the clinical practice, the relation between the input (in terms of inherent patterns apart from the given class label) and the output (decision targets/classes) is not that obvious, particularly when the correlation of signs, symptoms, test results of the patients could be the manifestation of multiple factors. The patterns discovered directly from the acquired data may have overlapping or functionally entwined AVAs as observed from our recent works [7, 9]. We call this pattern entanglement.

Hence, we present a new classification method, called Clinical Pattern Discovery and Disentanglement (cPDD), with novel capability to tackle this problem, particularly focused on the imbalanced class problem. The algorithm is briefly described in Fig. 1 by taking a relational dataset **R** says with N attributes as input.

Firstly, Attribute-Value Association Frequency Matrix (AVAFM) is constructed, where the Attribute-Value Association (AVA) is defined as the association between a pair of AVs (from different attributes). The AVAFM consists of the frequency of co-occurrences of all AV pairs within an entity from all entities in **R**. Then, to evaluate the statistical significance of each AVA, the frequency of co-occurrences in AVAFM is converted to a statistical measure known as adjusted statistical residual (SR) [6] accounting the deviation of that frequency from its default model, that is, the frequency of co-occurrences of the AV pairs is statistically independent, i.e., containing no correlated relation. Then, the new matrix is called AVA Statistical Residual Vector Space (SRV) each row of which represents an AV-vector with its coordinates representing the SR values of that AV associated with other AV's represented by the column vectors. The next step of cPDD is applying principal component decomposition (PCD) to decompose the SRV into different principal components (PCs) and re-project the projections of the AV-vectors on each PC to a new SRV, referred to as Re-projected SRV (RSRV). The AV-vectors with a new set of coordinates in the RSRV reflect the SR of AVAs captured by that PC. The PC and its RSRV together refer to an AVA Disentangled Space (DS). Since the number of DSs is as large as the number of AVs, cPDD only select a small set of DSs denoted by $\mathbf{DS}^* = \{ DS_i^* \}$ if the maximum SR in the RSRV of that DS exceeds a set statistical threshold (e.g., 1.44 in 85% confidence interval). As the AVs sharing statistically significant AVAs will form Attribute-Value Clusters (AV-Clusters) in a PC reflecting a group of strongly associating AVs. An AV Cluster is defined as a set of AVs such that each of which is associated with an AV of the other attribute in the cluster.

In traditional pattern discovery, to discover high-order patterns from the AVs of a dataset is complex since there is an exponential number of combinations of AVs as pattern candidates. cPDD discovers patterns from each of the small number of AV-Clusters from a small set $\mathbf{DS}^*$. Hence, it not only dramatically reduces the number of pattern candidates, but also separates patterns according to their orthogonal AVAs components revealing orthogonal functional characteristic (AVAs) in AV clusters [9, 10] and in subgroups of different DS*. Since the AV-clusters are coming from a disentangled source, the set of patterns discovered therein are relatively small with no or least overlapping and "either-or" cases among their AVs. Thus, cPDD significantly reduces the variance problem and relates patterns to more specific targets/groups. Unlike traditional Pattern Discovery (PD) methods which often produce an overwhelming number of entangled patterns, cPDD renders a much smaller succinct set of patterns associating with specific functionality from the disentangled sources for easy and direct interpretation. Furthermore, due to the reduction of the pattern-to-target variance, the patterns discovered from an uncorrelated AVA source environment will enhance prediction and classification, particularly effectively for data with imbalanced classes.

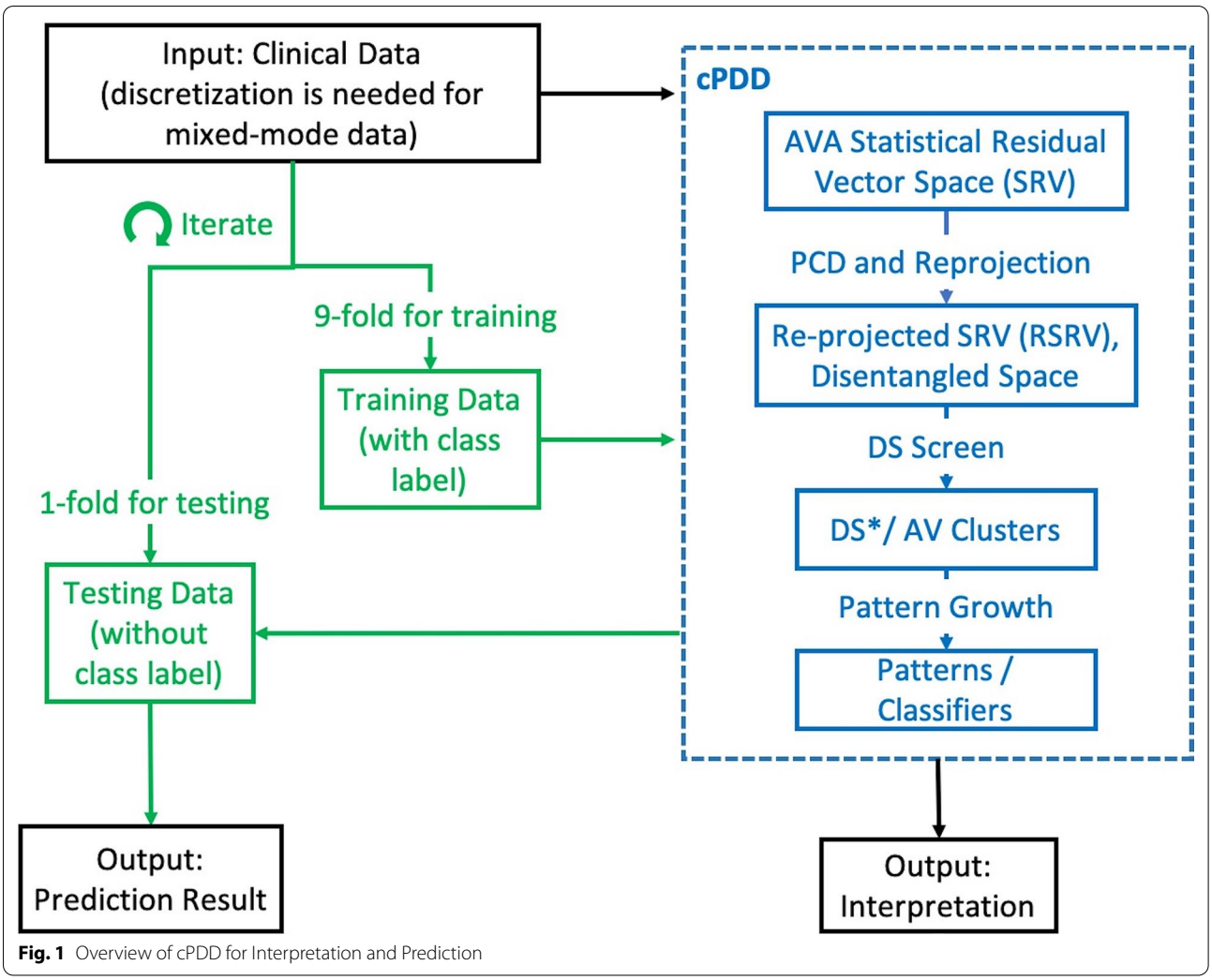


Fig. 1 Overview of cPDD for Interpretation and Prediction

Machine learning on clinical data analysis

Today, deep learning (DL) and frequent pattern mining are two commonly used methodologies for data analysis. However, in a more general healthcare setting where data analytics is based predominantly on clinically recorded numeral and descriptive data, the relation between the input (in terms of inherent patterns apart from the given class label) and the output (decision targets/classes) is not that obvious, particularly when the correlation of signs, symptoms, and test results of the patients could be the manifestation of multiple factors [2, 11]. Hence, this poses a challenge to DL in clinical application. Another concern is on the transparency and the assured accuracy [2, 11]. As for transparency, DL is generally considered as a black box [12]. Although ML methods like ensemble algorithms, such as Boosted SVM for imbalanced data (BSI), or Random Forest are good at prediction, their classification results are highly opaque and difficult for the clinicians to interpret [1]. Hence, to render transparency and interpretability, Decision Tree, Frequent Pattern Mining or Pattern Discovery were proposed. Since rules discovered by Decision Tree are guided by class labels, it is unlikely to discover class related AVAs between attributes when class labels are not available. Furthermore, as revealed in our recent work [7–9], AVAs discovered from relational data could be entangled due to multiple entwining functional characteristics inherent in the source environments. The patterns discovered using existing frequent pattern mining approaches based on the likelihood, weight of evidence [6], support, confidence or statistical residuals [5, 6], may have overlapping or functionally entwined AVA patterns inherent in the acquired data leading to overwhelming pattern number and redundancy, making explanation very difficult. Although extra pattern

clustering, pruning and summarization algorithms [13, 14] have been proposed and produced a smaller set of patterns/pattern clusters, yet the pattern entanglement problems have not been solved and the interpretation is not comprehensive and succinct.

The cPDD proposed in this paper has solved the fundamental pattern entanglement problem and met the clinical challenges posed above. It intends to provide clinicians with concise and robust clinical patterns discovered from the disentangled sources. The patterns are presented in a more succinct and interpretable form to reveal diagnostic characteristics of the patients and provide statistical support for prediction. Due to its ability of pattern disentanglement, patterns from minority class can be discovered in AVA Statistic Spaces (RSRVs) orthogonal to those of the majority classes.

cPDD extends our recent work [9] on AVA disentanglement to the discovery of statistically significant high-order patterns in AVA disentangled spaces. Its major contributions are three-fold.

i. The cPDD discovers and disentangles statistically significant high-order patterns to reveal the characteristics of different functional subgroups and/or classes in clinical data.
ii. It provides an explicit pattern representation for interpreting the characteristics of the dataset
iii. It uses the discovered patterns to classify entities in the dataset with high precision even when the class distributions are imbalanced.

Methods

In this section, we extend our previous work, Attribute-Value Association Discovery and Disentanglement Model (AVADD) [9, 10, 15], to cPDD to discover robust and succinct statistically significant high-order patterns and pattern clusters for interpreting and predicting clinical data with imbalanced classes. Table 1 gives an abbreviation of terms and Fig. 1 provides a schematic overview of cPDD.

First, we denote the input data as **R**, which contains N attributes, denoted as A = { A_1, A_2, ...A_N }, and each attribute (A_n) is denoted as $A_n = \{A_n^1, A_n^2, \ldots A_n^{I_n}\}$, where I_n represents the number of AVs of the nth attribute, A_n. The AVA for an AV pair (A_n^i, $A_{n'}^j$) is represented as $A_n^i \leftrightarrow A_{n'}^j$, which describes the association between "the *ith* value of the *nth* attribute" and "the *jth* value of the *n'th* attribute". Then, cPDD is implemented in the following five steps.

1. **Statistical Data Analysis:** The measurement that we used here is the same as that used for high-order pattern discovery [6] which uses a statistical method to evaluate the significance of the associations of an AV cluster. From here on, an AVA denotes an association between a pair of AVs (or called AV pair). First, the Frequency Matrix (FM) denoted by a $T \times T$ matrix of AVA relative frequencies between two AVs is constructed, where T is the total number distinct AVs in the table. Then the FM is turned into an Adjusted Statistical Residual Vector Space (SRV) to represent the statistical weights of all the AVA pairs obtained from **R**. The dimension of SRV is same as that of FM, and each item of SRV is denoted as a $\mathrm{SR}\left(A_n^i \leftrightarrow A_{n'}^j\right)$, which represents the adjusted residual between two AVs $\left(A_n^i \leftrightarrow A_{n'}^j\right)$. The value of $\mathrm{SR}\left(A_n^i \leftrightarrow A_{n'}^j\right)$ is calculated by Eq. (1)

Table 1 Notations and terminologies

AV	Attribute Value
AVA	Attribute Value Association
AV Cluster	Attribute Value Cluster
SR	Adjusted Statistical Residual for an AV pair
SRV	AVA Adjusted Statistical Residual Vector Space
PCD	Principal Component Decomposition
RSRV	Re-projected SRV
DS	Disentangled Space
DS*	Selected Disentangled Space, the selected set

$$SR\left(A_n^i \leftrightarrow A_{n'}^j\right) = r\left(A_n^i \leftrightarrow A_{n'}^j\right) v\left(A_n^i \leftrightarrow A_{n'}^j\right) \quad (1)$$

where $r\left(A_n^i \leftrightarrow A_{n'}^j\right)$ represents the standardized residual of the association.

$$r\left(A_n^i \leftrightarrow A_{n'}^j\right) = \frac{Occ\left(A_n^i \leftrightarrow A_{n'}^j\right) - Exp\left(A_n^i \leftrightarrow A_{n'}^j\right)}{\sqrt{Exp\left(A_n^i \leftrightarrow A_{n'}^j\right)}}.$$

$Occ\left(A_n^i \leftrightarrow A_{n'}^j\right)$ is the total number of co-occurrences for A_n^i *and* $A_{n'}^j$.

$Exp\left(A_n^i \leftrightarrow A_{n'}^j\right) = \frac{Occ(A_n^i)Occ\left(A_{n'}^j\right)}{M}$ is the expected frequency and M is the total number of entities.

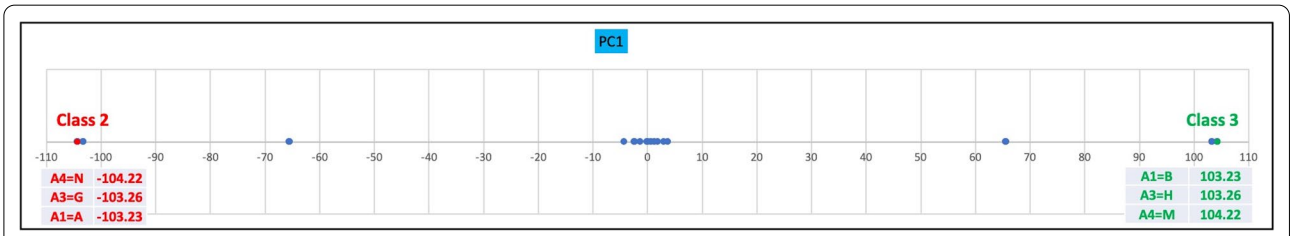
$v\left(A_n^i \leftrightarrow A_{n'}^j\right)$ represents the maximum likelihood estimate of the variance of $r\left(A_n^i \leftrightarrow A_{n'}^j\right)$,

and $v\left(A_n^i \leftrightarrow A_{n'}^j\right) = var\left(A_n^i \leftrightarrow A_{n'}^j\right) = 1 - \frac{Occ(A_n^i)}{M}\frac{Occ\left(A_{n'}^j\right)}{M}$

Therefore, SRV is an $T \times T$ matrix representing an adjusted standard residual (SR) Space [6] where T

represents the total number of distinct AVAs. Hence, SR accounts for the deviation of its observed frequency of occurrences against the expected frequency of occurrences if the AVs in the pair are statistically independent. Generally speaking, the significant associations can be selected according to the threshold obtained from the hypothesis test of statistically significant SR. For example, when the association's SR > 1.44, it can be treated as positively significant with an 85% confidence level (SR > 1.28 with 80% confidence level).

2. **Acquisition of AVA Disentangled Spaces**: For AVA disentanglement, Principal Component Decomposition (PCD) is applied to decompose the SRV into N PCs, denoted as PC = { PC_1, PC_2, ...PC_k...PC_N }, where PC_k is a set of projections of the AV vectors obtained from SRV after the PCD, and $PC_k = \{ PC_k(A_n^i) \mid n = 1, 2, \ldots N, i = 1, \ldots I_n \}$. We then re-project the projections of the AV-vectors captured in each PC to a new SRV with the same basis vectors and call it a Reprojected SRV (RSRV) corresponding to that PC. We then refer all the PCs and the their corresponding RSRVs = { $RSRV_1$, $RSRV_2$, ...$RSRV_k$, ...$RSRV_N$ } as the AVA disentangled spaces (DSs) where $RSRV_k$ is the re-projected result on PC_k via $RSRV_k = SRV \cdot PC_k \cdot PC_k^T$. Similar to SRV, each RSRV is an $I \times I$ matrix, and each row of a RSRV corresponding to an AV represents an AV-vector whose coordinates are the SR of that AV associating with other AVs represented by the column vectors in the RSRV. The coordinates of these AV vectors in the RSRV represent the SRs of the AVAs captured in the PCs. We refer to a PC with its RSRV as a Disentangled Space (DS). Figure 2 shows a DS (PC and RSRV) obtained from the synthetic dataset.
3. **Identification of functional sub-group (AV-Cluster)**: Since the number of DSs is as large as the number of AVs, we then devise a DS screening algorithm to select a small subset from DSs (denoted by **DS***) such that the maximum SR in its RSRV exceeds a statistical threshold (say 1.44 at confidence level of 85%). In the PC and RSRV of each DS*, often only one or two disjoint AV clusters are found. Each cluster may contain a few subgroups. Hence, the complexity of the PD process is greatly reduced. The criterion to form an AV cluster is that each AV in the cluster must be



RSRV1	Class=C3	A4=M	A3=H	A1=B	A6=F	A2=D	A8=O	A9=O	A7=Q	A9=P	A7=P	lass=C	A5=A	A5=B	A8=P	A8=Q	A7=O	A9=Q	A2=C	A6=J	A1=A	A3=G	A4=N	Class=C2
Class=C2	-33.83	-33.81	-33.50	-33.49	-21.28	-21.26	-1.20	-0.96	-0.59	-0.42	-0.19	-0.02	-0.02	0.02	0.45	0.76	0.78	1.38	21.26	21.28	33.49	33.50	33.81	33.84
A4=N	-33.80	-33.78	-33.47	-33.46	-21.26	-21.24	-1.20	-0.96	-0.59	-0.42	-0.19	-0.02	-0.02	0.02	0.45	0.75	0.78	1.38	21.24	21.27	33.46	33.47	33.78	33.81
A3=G	-33.72	-33.70	-33.39	-33.39	-21.30	-21.28	-1.42	-1.18	-0.82	-0.65	-0.42	-0.25	-0.25	-0.21	0.22	0.51	0.54	1.13	20.81	20.8	Class 2	2.93	33.24	33.27
A1=A	-33.25	-33.23	-32.92	-32.91	-20.83	-20.81	-0.95	-0.72	-0.35	-0.18	0.05	0.21	0.21	0.25	0.68	0.98	1.00	1.60	21.27	21.30	33.38	33.39	33.70	33.73
A6=J	-21.66	-21.65	-21.46	-21.45	-13.77	-13.76	-1.14	-0.99	-0.76	-0.65	-0.50	-0.40	-0.40	-0.37	-0.10	0.09	0.10	0.48	12.98	13.00	20.68	20.68	20.88	20.90
A2=C	-20.87	-20.85	-20.66	-20.65	-12.98	-12.97	-0.37	-0.22	0.01	0.12	0.27	0.37	0.37	0.40	0.67	0.86	0.87	1.25	13.74	13.76	21.43	21.43	21.63	21.65
A9=Q	-1.41	-1.41	-1.40	-1.40	-0.90	-0.90	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.03	-0.03	-0.01	0.00	0.00	0.02	0.83	0.83	1.33	1.33	1.34	1.34
A5=B	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.09	-1.09	-1.09	-1.09	-1.09	-1.09	-1.09	-1.09	-1.09	-1.09	-1.09	-1.09	-1.08	-1.08	-1.07	-1.07	-1.07	-1.07
A8=Q	-0.80	-0.79	-0.79	-0.79	-0.52	-0.51	-0.07	-0.06	-0.05	-0.05	-0.04	-0.04	-0.04	-0.04	-0.03	-0.02	-0.02	-0.01	0.43	0.43	0.71	0.71	0.71	0.71
A7=O	-0.76	-0.76	-0.75	-0.75	-0.47	-0.47	-0.01	0.00	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.04	0.04	0.05	0.51	0.51	0.79	0.79	0.80	0.80
A8=P	-0.44	-0.44	-0.43	-0.43	-0.27	-0.27	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.03	0.30	0.30	0.46	0.46	0.47	0.47
Class=C1	0.02	0.02	0.02	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.02	-0.02	-0.02	-0.02
A7=P	0.16	0.16	0.16	0.16	0.09	0.09	-0.02	-0.02	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.04	-0.15	-0.15	-0.22	-0.22	-0.22	-0.22
A9=P	0.45	0.44	0.44	0.44	0.29	0.29	0.04	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.01	-0.24	-0.24	-0.39	-0.39	-0.39	-0.39
A7=Q	0.60	0.60	0.59	0.59	0.38	0.38	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.00	-0.01	-0.01	-0.02	-0.36	-0.36	-0.58	-0.58	-0.58	-0.58
A9=O	0.97	0.97	0.96	0.96	0.61	0.61	0.04	0.03	0.02	0.02	0.01	0.01	0.01	0.01	-0.01	-0.01	-0.02	-0.03	-0.60	-0.60	-0.94	-0.94	-0.95	-0.95
A5=A	1.11	1.11	1.11	1.11	1.11	1.11	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.08	1.08	1.07	1.07	1.07	1.07
A8=O	1.22	1.22	1.21	1.21	0.78	0.78	0.07	0.06	0.05	0.04	0.03	0.03	0.03	0.03	0.01	0.00	0.00	-0.02	-0.73	-0.73	-1.16	-1.16	-1.17	-1.17
A2=D	20.87	20.85	20.66	20.65	12.98	12.97	0.37	0.22	-0.01	-0.12	-0.27	-0.37	-0.37	-0.40	-0.67	-0.86	-0.87	-1.25	-13.74	-13.76	-21.43	-21.43	-21.63	-21.65
A6=F	21.66	21.65	21.46	21.45	13.77	13.76	1.14	0.99	0.76	0.65	0.50	0.40	0.40	0.37	0.10	-0.09	-0.10	-0.48	-12.98	-13.00	-20.68	-20.68	-20.88	-20.90
A1=B	33.25	33.23	Class 3	1	20.83	20.81	0.95	0.72	0.35	0.18	-0.05	-0.21	-0.21	-0.25	-0.68	-0.98	-1.00	-1.60	-21.27	-21.30	-33.38	-33.39	-33.70	-33.73
A3=H	33.72	33.70		9	21.30	21.28	1.42	1.18	0.82	0.65	0.42	0.25	0.25	0.21	-0.22	-0.51	-0.54	-1.13	-20.81	-20.84	-32.92	-32.93	-33.24	-33.27
A4=M	33.80	33.78	33.47	33.46	21.26	21.24	1.20	0.96	0.59	0.42	0.19	0.02	0.02	-0.02	-0.45	-0.75	-0.78	-1.38	-21.24	-21.27	-33.46	-33.47	-33.78	-33.81
Class=C3	33.82	33.80	33.49	33.48	21.28	21.25	1.20	0.96	0.59	0.42	0.19	0.02	0.02	-0.02	-0.45	-0.75	-0.78	-1.38	-21.25	-21.28	-33.48	-33.49	-33.80	-33.83

Fig. 2 An illustration of DS* with two AV clusters in the first PC (PC1) using synthetic data. As displayed in PC1, two distinct clusters far from the centre represent two strongly AVA groups corresponding to class 2 and class 3 with large eigenvalue. The SR of their corresponding AVA among the AV pairs in each cluster are shown by the yellow shaded cells in its corresponding Re-projected SRV (RSRV). This indicates that the AVAs of class 2 and class 3 are disentangled and grouped in the first PC and its corresponding RSRV. The green shaded cells in RSRV denote AV pairs with negative statistical significance (i.e., very unlikely to occur)

in a significant AVA with other AV in the cluster. In the RSRV (Fig. 2), the cells with yellow and green shade show the AV pairs with positive and negative statistical significance respectively.

4. **Pattern Discovery**: High-order patterns can be discovered through identifying pattern candidates through an AV cluster identification and growing process. Formally, we denote a high-order pattern as P_j which consists of a subset of AVs with size ≥ 2. We use the adjusted residual [2] derived from the frequency of co-occurrences of P_j used in the hypothesis test to assess whether P_j is a statistically significant pattern. In order to keep the discovered patterns non-redundant, we only accept delta-closed patterns [16, 17] in the pattern discovery process. There might be more than one pattern identified in the AV cluster. We treat the union of the AVs making up patterns in one AV cluster or in one functional sub-group as the summarized super pattern. All patterns discovered by cPDD are listed as the comprehensive patterns.
5. **Interpretation and Prediction**: The AVs in each AV cluster/subgroup making up a summarized pattern pertaining to a designated class/group. In all our experiments, due to AVA disentanglement, the summarized patterns contain no or very few "either-or AVs" within the pattern. Hence, the summarized pattern is more succinct and easier to interpret. The high-order patterns in the comprehensive set can provide all the detailed patterns for interpretation and linkage to individuals and groups. Since the number of candidate AVs are few in the output of cPDD, so the number of patterns discovered in each DS* is extremely small. This is significantly different from traditional PD. For class prediction when class labels are given, we can discover the disentangled patterns associating with class labels from the training data. In testing, we apply the discovered summarized patterns associated with each specific class to predict whether the entity for testing belongs to that class. Let (P_j, C) represents a summarized pattern P_j associated with class label C, and E_i represent the entity needed to be predicted. Based on the mutual information in statistical information theory, we can use the weight of evidence [18, 19] of all the AVs in the summarized patterns to determine whether the class label for E_i, $C(E_i)$, will have higher weight than predicting it as pertaining to other classes.

Results and discussion

In this study, we conducted experiments both on the synthetic and the clinical dataset with imbalanced classes.

Materials

Dataset 1: synthetic dataset

To show the capability of cPDD in interpreting an imbalance dataset, a synthetic experiment was designed and conducted. We generated stochastically a 2100×10 matrix with the first column as the class label and others as attributes with character values stochastically generated from a uniform distribution. This represents a random relational dataset with attributes independent to each other. We then embedded patterns of three different classes C_1, C_2, and C_3 for the first 6 attributes. We use A1A, A2C, for example, to respectively represent character value A and C for Attribute A1 and A2. The patterns implanted in the data are summarized in Table 2. Note that A1A and A2C are entangled (overlapping) for C_1 and C_2; A3H and A4M are entangled in C_1 and C_3; A5B and A6J are entangled in C_2 and C_3. For the last three attributes, we put in randomly selected characters from {"O", "P", "Q"} and for the 10th attribute we randomly embedded characters used for the three classes. Moreover, this synthetic Dataset was implemented as one with imbalanced class distribution with 1000 entities pertaining to C_2 and C_3 each, and 100 entities pertaining to C_1.

Dataset 2: thoracic dataset

The thoracic dataset describes the surgical risk originally collected at Wroclaw Thoracic Surgery Centre for patients who underwent major lung resections for primary lung cancer in the years 2007-2011 [20]. The attributes included are given in Fig. 3. This public dataset is provided after feature selection and elimination of missing values. It is composed of 470 samples with 16 pre-operative attributes after feature selection. The target attribute (class label) is Risk. Risk = T if the patient died. In this dataset, the class distribution is

Table 2 Synthetic dataset with embedded entangled patterns

Classes	Attribute Values are Significant Associated with Class Label
C1	A1A, A2C, A3H, A4M/N, A5A, A6F
C2	A1A, A2C/D, A3G, A4N, A5B, A6J
C3	A1B, A2D, A3H, A4M, A5B, A6F/J

1. DGN	Diagnosis - specific combination of ICD-10 codes for primary and secondary as well multiple tumours if any (DGN3,DGN2,DGN4,DGN6,DGN5,DGN8,DGN1)
2. PRE4	Forced vital capacity - FVC (numeric)
3. PRE5	Volume that has been exhaled at the end of the first second of forced expiration - FEV1 (numeric)
4. PRE6	Performance status - Zubrod scale (PRZ2,PRZ1,PRZ0)
5. PRE7	Pain before surgery (T,F)
6. PRE8	Haemoptysis before surgery (T,F)
7. PRE9	Dyspnoea before surgery (T,F)
8. PRE10	Cough before surgery (T,F)
9. PRE11	Weakness before surgery (T,F)
10. PRE14	T in clinical TNM - size of the original tumour, from OC11 (smallest) to OC14 (largest) (OC11,OC14,OC12,OC13)
11. PRE17	Type 2 DM - diabetes mellitus (T,F)
12. PRE19	MI up to 6 months (T,F)
13. PRE25	PAD - peripheral arterial diseases (T,F)
14. PRE30	Smoking (T,F)
15. PRE32	Asthma (T,F)
16. AGE	Age at surgery (numeric)
17. Risk1Y	**1 year survival period - (T)rue value if died (T,F)**

Fig. 3 Attribute Description of Thoracic Dataset

imbalanced with 70 cases being Risk = T and 400 cases being Risk = F. To simulate the target scenario without requiring much tweaking, the numeric attributes PRE4, PRE5 and age were removed.

Analysis I – discovery and display of explicit patterns for explanation

In Analysis I, we compared the discovered patterns obtained in cPDD, Apriori [21] (a typical frequent pattern mining method) and a high-order pattern discovery method for discrete-value data (HOPD) [6] which was our early work closely resembling the PD reported in [9, 10]. Figures 4 and 5 show the pattern discovery result of cPDD on the Synthetic and Thoracic data respectively. Figure 6 presents the comparison results of all these three methods.

As shown in Fig. 4, a small set of AV-Clusters was discovered from the synthetic dataset. Figure 4a displays the union of all the comprehensive patterns (Fig. 4b) and can be considered as the summarized pattern. The summarized pattern in each subgroup of a DS* consists the union of all the detailed patterns discovered in that subgroup of the DS*. While the summarized pattern gives a high-level interpretation of the AVs with significant AVAs in a subgroup, the detailed patterns encompass comprehensively all the significant patterns discovered in that subgroup with details and statistical support. In the like manner, the summarized patterns discovered form the Thoracic dataset are given in Fig. 5a and some samples of comprehensive set of patterns that are associated with class labels are displayed in Fig. 5b.

a

DS 1		Class	A1	A2	A3	A4	A5	A6	A7	A8	A9
AV Cluster	1	C2	A	C	G	N		J			
	2	C3	B	D	H	M		F			
DS 2		**Class**	A1	A2	A3	A4	A5	A6	A7	A8	A9
AV Cluster	1	C1	A	C	H		A	F			
	2	C2/C3	B	D	G		B	J			

b

		Detailed Patterns									
DS1	**Residual**	**Class**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
AV Cluster 1	98.57	C2	A		G	N		J			
	94.59	C2	A	C	G	N		J			
AV Cluster 2	98.59	C3	B	D	H	M					
	94.59	C3	B	D	H	M		F			
DS2	**Residual**	**Class**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
AV Cluster 1	304.36	C1	A	C	H		A	F			
AV Cluster 2	40.81	C3	B	D			B				
	40.81	C2			G		B	J			
	18.7	C3	B	D			B	J			
	18.7	C2		D	G		B	J			

c

Implanted Patterns		Pattern Discovery by cPDD		
		Residual	**Class**	**Disentangled Space and AV Cluster**
P1	A1A, A2C, A3H, A4M/N, A5A, A6F	304.36	C1	DS2; AV Cluster 1
P2	A1A, A2C, A3G, A4N, A5B, A6J	98.57	C2	DS1; AV Cluster 1
	A1A, A2C, A3G, A4N, A5B, A6J	94.59	C2	DS1; AV Cluster 1
	A1A, A2D, A3G, A4N, A5B, A6J	*18.7*	C2	DS2; AV Cluster 2
P3	A1B, A2D, A3H, A4M, A5B, A6F/J	98.59	C3	DS1; AV Cluster 2
	A1B, A2D, A3H, A4M, A5B, A6F	94.59	C3	DS1; AV Cluster 3
	A1B, A2D, A3H, A4M, A5B, A6F/	40.81	C3	DS2; AV Cluster 4
	A1B, A2D, A3H, A4M, A5B, A6FJ	18.7	C3	DS2; AV Cluster 5

Fig. 4 cPDD Pattern Discovery Result from Synthetic Dataset. **a** In the First and the Second DS* (DS 1 and DS 2) two AV-clusters were discovered. **b** Detailed Patterns associated with different classes were discovered from the above two AV-Clusters. **c** Implanted patterns showing close correspondence to the patterns obtained in cPDD's output in (**a**) and (**b**)

Figure 6 displayed the PD results obtained from cPDD and Apriori with two different fine-tuned sets of support and confidence on the synthetic dataset (Fig. 6a) and the Thoracic dataset (Fig. 6b). While cPDD discovers 4 summarized patterns and 9 detailed patterns from the synthetic dataset, the number of patterns discovered by Apriori and HOPD are overwhelming. For the thoracic dataset, similar comparative results for are shown in Fig. 6.

Furthermore, when comparing the implanted patterns with the patterns cPDD discovered (Fig. 4c), cPDD reveals all patterns with correct class labels in disentangled spaces except one in P2 as it has (A2D, A3G, A5B, A6J). Although this pattern is the same as the implanted patterns, yet it shares the sub-pattern (A2D, A5B, A6J) of P3 in C3 indicating the entanglement in the original data. Figure 4c shows high SR for the implanted patterns assigned with the correct classes and low SR for the entangled cases. For both Apriori or HOPD, they

a

DS 1		Risk	Diagnosis	PRE6	PRE7	PRE8	PRE9	PRE10	PRE11	PRE14	PRE17	PRE19	PRE25	PRE30	PRE32
AV Cluster	1	T	DGN2	PRZ1/PRZ2		T	T	T	T	OC14/OC13			T	T	
	2	F		PRZ0		F	F	F	F	OC11			F	F	
DS 2		**Risk**	**Diagnosis**	**PRE6**	**PRE7**	**PRE8**	**PRE9**	**PRE10**	**PRE11**	**PRE14**	**PRE17**	**PRE19**	**PRE25**	**PRE30**	**PRE32**
AV Cluster	1	T	DGN5		T	T	T		F	OC13				F	
	2	F	DGN3		F	F	F		T	OC11				T	

b

		Detailed Patterns													
DS1	**Residual**	Risk	Diagnosis	PRE6	PRE7	PRE8	PRE9	PRE10	PRE11	PRE14	PRE17	PRE19	PRE25	PRE30	PRE32
AV Cluster 1	1.89	T								OC14				T	
	2.33	T						T		OC14					
	1.66	T		PRZ1						OC14					
	1.71	T		PRZ1				T		OC14					
	1.27	T		PRZ1			T	T							
	1.56	T		PRZ1				T	T						
	6.35	T	DGN2					T		OC14					
	2.12	T						T		OC14				T	
	3.05	T		PRZ1				T						T	
	2.1	T					T	T						T	
	2.22	T	DGN2					T						T	
	2.87	T						T	T					T	
	1.89	T	DGN2						T					T	
	3.1	T				T		T	T					T	
	7.38	T		PRZ2				T	T					T	
	1.81	T	DGN2	PRZ1				T						T	
	2.95	T		PRZ1		T		T	T					T	
AV Cluster 2	3.4	F		PRZ0			F		F						
	2.83	F		PRZ0		F				OC11					
	2.65	F					F	F	F						
	4.72	F		PRZ0		F	F		F	OC11					
	4	F		PRZ0		F	F		F				F		
	16.19	F		PRZ0		F	F	F	F						
	15.07	F		PRZ0			F	F	F				F		
	4.53	F				F	F	F	F	OC11					
	3.38	F				F	F	F	F				F		
	1.24	F				F	F		F				F	F	
	2	F				F	F		F	OC11			F	F	
	13.95	F		PRZ0		F	F	F	F	OC11			F		
	11.93	F		PRZ0		F	F	F	F				F	F	
	10.07	F		PRZ0		F	F	F	F	OC11			F	F	
DS2	**Residual**	**Risk**	**Diagnosis**	**PRE6**	**PRE7**	**PRE8**	**PRE9**	**PRE10**	**PRE11**	**PRE14**	**PRE17**	**PRE19**	**PRE25**	**PRE30**	**PRE32**
AV Cluster 1	1.6	T						T	T					T	
AV Cluster 2	1.68	F			F	F	F							T	
	1.27	F			F		F		T					T	
	2.6	F	DGN3		F	F	F							T	
	2.91	F	DGN3		F	F	F			OC11				T	

Fig. 5 Pattern Discovery Result of Thoracic Dataset using cPDD **a** In both First and the Second Disentangled Space, two AV Clusters corresponding to Risk=T and RISK=F were discovered **b** Detailed Patterns discovered from the above two AV Clusters. Note that no AV of attributes PRE17, PRE19 and PRE32 are associated with any significant patterns. They were dismissed autonomously without undergoing feature engineering

discovered a large number of patterns where most of them are redundant and overlapping. While some of them are associated with class labels, others are with the noise columns A7, A8 and A9 as well.

In addition, from the pattern discovery result on the Thoracic dataset, we observed similar phenomena if we replace Figs. 4 and 6a with Figs. 5 and 6b respectively. Figure 5a gives four AV-Clusters, two in each AVA disentangled Space (DS1 and DS2). Each AV-cluster contains the interpretable union of all the patterns discovered in different subgroups (Fig. 5a). A subset of the detailed patterns forming each union pattern is displayed in Fig. 5b.

Figure 6b shows the comparison results for the Thoracic data. First, the number of patterns obtained by Apriori and HOPD are both large. And it is difficult to interpret the pattern outcomes relevant to the problem when the number of patterns is large with considerable redundant and overlapping patterns. Second, Apriori outputs the patterns from datasets only if the class labels are given. HOPD can output all the patterns discovered among the growing set of the candidate

a

Synthetic Data	Apriori		HOPD	cPDD	
	Sup=10%; Con=20%	Sup=10%; Con=10%		Summarized Patterns	Detailed Pattenrs
Patterns Associate with Imbalanced Class	No	Yes	No	Yes	Yes
# of Patterns	946	962	770	4	9

b

Throcic Data	Apriori		HOPD	cPDD	
	Sup=10%; Con=30%	Sup=10%; Con=20%		Summarized Patterns	Detailed Pattenrs
Patterns Associate with Imbalanced Class	No	Yes	Yes	Yes	Yes
# of Patterns	18071	18363	9513	4	36

Fig. 6 **a** Comparison of Pattern Discovery Result on Synthetic Dataset **b** Comparison of Pattern Discovery Result on Thoracic Dataset. The number of patterns discovered by cPDD is drastically reduced to enable succinct interpretation and the construction of a useful knowledge base

patterns without knowing class labels, but the number of high order patterns produced are overwhelming. For a dataset **R** with m attributes, there are an exponential number of AV combinations being considered as pattern candidates. So, the number of patterns outputted by HOPD is huge. It is surprising to note that the highest order of patterns discovered for Thoracic Dataset by cPDD is 8 (Fig. 5b), yet the number of comprehensive patterns discovered is only 36 (Fig. 6b). This is really beyond what humans can grasp.

When we examined whether other algorithms could discover the patterns associated with the minority class, we found that the results of Apriori depend on the set value of the threshold, support, and confidence. When the threshold is low, more patterns are discovered which may cover those in the minority class, but the number of patterns is huge (Fig. 6b). When the set threshold was set high, patterns in the rare class were not discovered. As for HOPD, it discovered a large number of patterns that contain those of the rare classes. However, cPDD discovered a much smaller number of summarized and detailed patterns succinctly, including those from the rare class.

In summary, this experimental result shows that cPDD is able to discover fewer patterns with specific association to the classes in support of easy and feasible interpretation. Furthermore, even with few patterns, it is able to represent succinct, comprehensive (as exemplified in the synthetic case) and statistical/functional characteristics of all classes given, even when the class distribution is imbalanced. With the capability to render direct interpretation of a small, succinct and reliable set of patterns discovered from distinct sources without the reliance of explicit a priori knowledge and a posteriori processing, cPDD is a novel approach of Explainable AI (XAI) [22, 23] quite different from the existing model-based ML approach.

Analysis II – prediction on imbalanced dataset

In Analysis II, we focus on the prediction of diagnostic outcomes of the Thoracic dataset with imbalance class distribution. We first report the testing results on the original thoracic dataset, then we cover the extended experiment results with sampling data.

Comparison result on original data

For the imbalanced class problem, since the correct prediction of the majority classes will overwhelm that of the minority classes, the prediction performance should not be evaluated based on the average accuracy [24]. Hence, in this study, the prediction results are evaluated by the F1-Score [25] calculated by Precision and Recall (or called sensitivity and specificity), Geometric mean of Precision and Recall (G-mean) [1] respectively for predicting the minority target.

The F1-Score for the minority class C_m, denoted as $F1(C_m)$, can be calculated from the *Precision* (C_m) and the *Recall* (C_m) by Eq. (2).

$$F1(C_m) = \frac{2 * Recall\ (C_m) * Precision(C_m)}{Recall\ (C_m) + Precision(C_m)} \tag{2}$$

where $Precision\ (C_m) = \frac{TP}{(TP+FP)}$, $Recall\ (C_m) = \frac{TP}{(TP+FN)}$, *TP* represents True Positive; *FP* represents False Positive, and *FN* represents False Negative when considering the minority class label as the target. Thus, according to the definition, F1-score = 0 if the number of true positive $TP = 0$.

The G-mean for the minority class, denoted as G-mean(C_m), is calculated from the *Precision* (C_m) and the *Recall* (C_m) by Eq. (3).

$$G - mean(C_m) = \sqrt{Recall\ (C_m) * Precision(C_m)} \tag{3}$$

We do not show the comparison result of the Precision or Recall of minority class since for the imbalanced data problem, if all the cases are detected as majority class target or minority class target, they may have extremely high precision or recall. Hence, the average F1-score or G-mean calculated by both Precision and Recall are obtained from 20-time-10-fold cross-validation used for performance evaluation.

Besides, considering both majority and minority decision in the comparison result, we list the average accuracy to show how misleading these results could be as they are unreliable measures for classification performance evaluation for the case with imbalanced classes. Compared to Accuracy and F1 score, MCC renders a more reliable statistical rate which produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (TF, FN, TN and FP) [26]. The value of MCC is from −1 to +1, where +1 represents a perfect prediction, 0 an average random prediction and −1 an inverse prediction. Hence, wwe also list the Matthews correlation coefficient (MCC) [26] (Eq. (4)) and Balanced Accuracy (Balance Acc.) [27] to show the accuracy after the balanced results are obtained.

$$MCC = \frac{\frac{TP}{N} - S * P}{\sqrt{PS(1-S)(1-P)}}$$
$$\text{where } N = TN + TP + FN + FP,\ S = \frac{TP+FN}{N},\ P = \frac{TP+FP}{N} \tag{4}$$

In this study, cPDD was compared with Logistic Regression (LR), Naïve Bayes (NB), and Decision Tree (CART). All of the above algorithms were implemented under default parameters using the Python machine learning package, scikit-learn 0.23.2 [28]. The comparison results are given in Table 3 and Fig. 7. All results are listed as *mean* ± *variance*.

Table 3 Comparison result from 20-time-10-fold cross-validation using different classification algorithm

Original Thoracic Data	LR	CART	NB	cPDD
F1-Score(T)	0.01 ± 0.00	0.19 ± 0.03	0.24 ± 0.01	0.33 ± 0.01
G-mean(T)	0.01 ± 0.01	0.20 ± 0.03	0.36 ± 0.01	0.38 ± 0.01
Avg. F1	0.01 ± 0.01	0.19 ± 0.03	0.24 ± 0.01	0.33 ± 0.01
Avg. Acc	0.84 ± 0.84	0.82 ± 0.01	0.15 ± 0.00	0.59 ± 0.01
Balanced Acc.	0.50 ± 0.00	0.54 ± 0.01	0.50 ± 0.00	0.62 ± 0.01
MCC	−0.01 ± 0.00	0.11 ± 0.03	0.01 ± 0.00	0.18 ± 0.01

F1-Score(T) Average F1-Score on Risk = T, *G-means(T)* Average G-mean on Risk = T, *Avg. F1* Average F1-Score for both classes (Risk = T and Risk = F), *Avg. Acc* Average of prediction accuracy on the whole dataset

In Table 3, LR shows poor prediction performance, resulting in 0.01 ± 0.00 on F1-Score and 0.01 ± 0.01 on G-mean. CART achieved a slightly better performance on F1-Score with 0.19 ± 0.03 and on G-mean with 0.20 ± 0.03 since the weighted samples were used for optimizing CART [28]. Naive Bayes was less influenced as the target proportion could be used as the prior information in training. Finally, we found that cPDD achieved the best performance on F1-Score with 0.33 ± 0.01 and G-means with 0.38 ± 0.01. In addition, for this set of imbalanced data, Balance Acc. would be better than regular accuracy [27]. Without sampling, both LR and NB shows the balanced accuracy as 0.5 which is same with randomly selection, even though NB shows better performance of classification for minority class.

Similarly, as Table 3 shows, CART obtained a slightly higher value on Balanced Acc and MCC. However, cPDD obtained the highest results for both Balanced Acc. with 0.62 + =0.01 and MCC with 0.18 + =0.01, as well as all the other scores except for Avg. Acc. which is not that meaningful for the imbalance cases.

In summary, cPDD has achieved the best result no matter whether it is applying on the entire dataset or only on the minority classes. This indicates that cPDD is more robust and reliable.

Comparison result on sampling data

To reduce the inaccuracy of such kind of biased classification, researchers usually use undersampling and oversampling methods to balance the samples of the training data [29]. Therefore, both random oversampling [30] and random undersampling [31] have been applied to the dataset before training and predicting.

Random oversampling Random oversampling [30] duplicates records randomly from the minority class and adds them to the training dataset. However, oversampling may result in overfitting towards the minority class

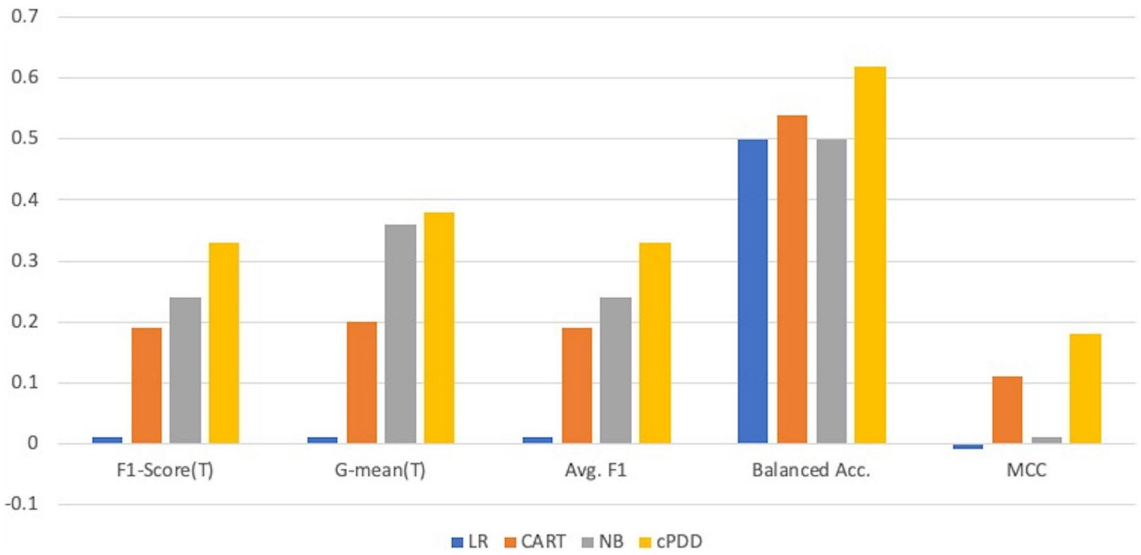


Fig. 7 Comparison Result from 20-time-10-fold cross-validation on Thoracic Dataset

Table 4 Comparison result from 20-time-10-fold cross validation with different sampling strategies

	LR	CART	NB	cPDD
Over Sampling				
F1-Score(T)	0.31 ± 0.02	0.20 ± 0.02	0.26 ± 0.01	0.37 ± 0.01
G-mean(T)	0.34 ± 0.02	0.22 ± 0.22	0.36 ± 0.00	0.40 ± 0.01
Avg. F1	0.31 ± 0.02	0.20 ± 0.20	0.26 ± 0.01	0.59 ± 0.01
Balanced Acc.	0.61 ± 0.01	0.52 ± 0.01	0.50 ± 0.00	0.61 ± 0.01
MCC	0.17 ± 0.03	0.04 ± 0.02	0.01 ± 0.00	0.25 ± 0.01
Under Sampling				
F1-Score(T)	0.30 ± 0.01	0.27 ± 0.02	0.25 ± 0.01	0.34 ± 0.02
G-mean(T)	0.34 ± 0.01	0.30 ± 0.02	0.35 ± 0.02	0.41 ± 0.03
Avg. F1	0.30 ± 0.01	0.27 ± 0.02	0.25 ± 0.01	0.63 ± 0.02
Balanced Acc.	0.59 ± 0.01	0.57 ± 0.01	0.54 ± 0.01	0.61 ± 0.02
MCC	0.13 ± 0.02	0.11 ± 0.02	0.08 ± 0.02	0.20 ± 0.08

F1-Score(T) Average testing F1-Score on Risk = T, *G-means(T)* Average testing G-mean on Risk = T, *Avg. F1* Average testing F1-Score for both classes (Risk = T and Risk = F)

samples especially for higher over-sampling rates [29, 32]. To implement the random oversampling algorithm, we used the imbalanced-learn Python library [33].

In this experiment, we separated the training data and testing data first, and then applied random oversampling on the training data. The original training dataset would include 423 records with ~63 records taken from the minority class and ~360 from the majority class. After applying the above random oversampling method, the number of training records was increased to 720. To keep consistent with experiments on original data, we used the original 47 test samples (10% of the original entire set) with imbalanced classes as the test set. We then applied all classification methods for training and testing on this set of data. Since we have shown that the average accuracy should not be used as a reasonable approach for prediction evaluation for imbalanced dataset, we listed in Table 4 only the average F1-score and G-mean for minority class and the average F1-score for the entire dataset.

As the results show, cPDD still achieved superior prediction results, since it can handle imbalanced dataset effectively. The oversampling strategy did not change the result too much for cPDD as its F1-score and G-mean increased respectively to (0.37 ± 0.01) and (0.40 ± 0.01) from (0.33 ± 0.01) and (0.38 ± 0.01) in Table 3. Similarly, both CART and NB would be less influenced by imbalanced data as the results only slightly increased with F1-Score 0.20 ± 0.02 and G-mean 0.22 ± 0.22 for CART, and almost kept the same for NB with F1-Score 0.26 ± 0.01 and G-mean 0.36 ± 0.00. However, since Logistic Regression is not designed for imbalanced datasets, so after oversampling, the dataset became a balanced dataset, and the performance of LR improved considerably with F1-score (0.31 ± 0.02) and G-mean (0.34 ± 0.02) for the minority class.

Random undersampling Similar to the random oversampling, random undersampling [31] randomly deletes records from the majority class. The process is repeated until the training dataset becomes a balanced dataset. The same Python library [33] was used for implementation. And the average F1-Score, G-mean for minority

class and the average F1-Score, MCC and Balance Acc. for the entire testing data were used for evaluation.

In this experiment, after applying the random undersampling algorithm, the number of training dataset (423 records) was reduced to 126 since the size of majority class was reduced. Then the same classification methods were applied to the same testing dataset.

As Table 4 shows, cPDD still achieved superior prediction results with F1-score (0.34 ± 0.02) and G-mean (0.41 ± 0.02). The results were also improved for CART with F1-Score in 0.27 ± 0.02 and G-mean in 0.30 ± 0.02 whereas the results of NB were least influenced by undersampling with F1-Score in 0.25 ± 0.01 and G-mean in 0.31 ± 0.02 while those of Logistic Regression improved with F1-score (0.30 ± 0.01) and G-mean (0.34 ± 0.01) for the minority class.

Besides, the comparison plots of F1-Score and G-mean are shown in Fig. 8. As the Fig. 8 shows, both cPDD and NB were less influenced by imbalanced data because both of them are "probabilistic classifiers" using statistical theory. cPDD could achieve better performance because it can discover even the hidden patterns. Logistic Regression cannot handle imbalanced data, so after samplings, the performance of LR improved considerably. Comparing between different sampling strategies, we found that CART could achieve better results when undersampling strategy was applied, and LR could obtain better results when oversampling strategy was applied.

In summary, without sampling, cPDD showed robust prediction performance in comparison with all other methods. And for all sampling strategies, cPDD still performed best. Thus, no matter whether the data is balanced or imbalanced, cPDD can handle it robustly and steadily.

Conclusion

As a pattern discovery method on imbalanced synthetic and Thoracic data, cPDD renders a much smaller succinct set of explicit class-associated patterns for better interpretation, and superior prediction since it uses disentangled patterns which are more specific and distinct to the classes. The results it obtains are statistically robust with comprehensive coverage of succinct, concise, precise, displayable and less redundant representations for experts' interpretation. cPDD also overcomes the limitations of lack of transparency [11] as well as the problem of imbalanced class [2, 3, 11, 34]. As a clinical data analysis tool on relational data, it has a significant advantage over the 'black box' ML algorithms since its output of is both transparent and interpretable, the two major challenges of interpretability and applicability [22] confronting ML on relational data today. The experimental result on synthetic and clinical data with high imbalanced class ratios shows that cPDD does have a superior prediction and interpretability performance for minority targets. cPDD brings explainable AI to clinical experts to enhance their insight and understanding with statistical and rational accountability. Hence, it will have great potential to enhance ML and Explainable AI [22, 23].

In our future work, cPDD will be developed to apply to unstructured data (e.g., text and sequences) [7, 35] by extracting patterns directly from them as shown in our early work [36]. Moreover, for performance improvement, parallel computing strategy will be introduced to handle bigger data and further speed up the computational time.

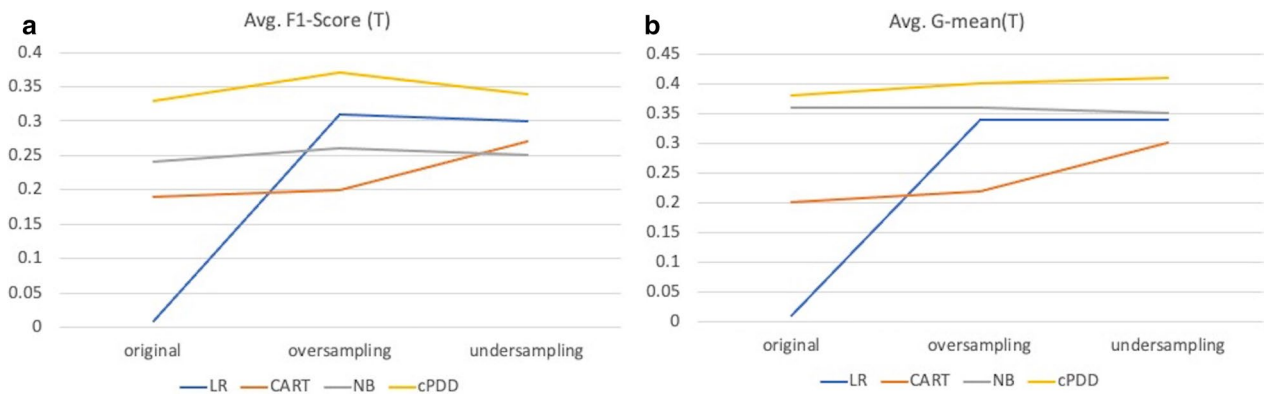


Fig. 8 Comparison Results from 20-time-10-fold cross validation on Risk = T with different sampling strategies. **a** Comparison of average testing F1-Score Using Various Classification Algorithms **b** Comparison of average testing G-mean Using Various Classification Algorithms

Abbreviations
AV: Attribute value; AVA: Attribute Value Association; AV Cluster: Attribute Value Cluster; SR: Adjusted Statistical Residual for an AV pair; SRV: AVA Adjusted Statistical Residual Vector Space; PCD: Principal component decomposition; RSRV: Re-projected SRV; DS: Disentangled Space; DS*: Selected Disentangled Space, the selected set.

Acknowledgements
Not applicable.

Authors' contributions
PZ and AW directed and designed the study, performed the clinical analyses and prepared the manuscript. PZ implemented the algorithm and performed the statistical analyses. All authors read and approved the final manuscript.

Funding
Publication cost is funded by NSERC Discovery Grant (xxxxx 50503-10275 500). This funding source had no role in the design of this study and will not have any role during its execution, analyses, interpretation of the data, or decision to submit results

Availability of data and materials
The thoracic dataset is available at from the University of California Irvine Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data.

Ethics approval and consent to participate
Not required as the datasets were published retrospective datasets, where the corresponding approval and consent were handled specifically

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 8 May 2020 Accepted: 30 November 2020
Published online: 09 January 2021

References

1. Chan T, Li Y, Chiau C, Zhu J, Jiang J, Huo Y. Imbalanced target prediction with pattern discovery on clinical data repositories. BMC Med Inform Decis Mak. 2017;17(1):47.
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56.
3. Aggarwal C, Sathe S. Bias reduction in outlier ensembles: the guessing game. In: Outlier ensembles: Springer; 2017.
4. Naulaerts S, Meysman P, Bittremieux W, Vu TN, Vanden Berghe W, Goethals B, Laukens K. A primer to frequent itemset mining for bioinformatics. Brief Bioinform. 2015;16(2):216–31.
5. Aggarwal C, Bhuiyan M, Hasan M (2014) Frequent pattern mining algorithms: a survey. In: Aggarwal C, Han J, editors. Frequent pattern mining. Cham: Springer. https://doi.org/10.1007/978-3-319-07821-2_2.
6. Wong AK, Wang Y. High-order pattern discovery from discrete-valued data. IEEE Trans Knowl Syst. 1997;9(6):877–93.
7. Zhou P-Y, Lee AE, Sze-To A, Wong AK. Revealing subtle functional subgroups in class A scavenger receptors by pattern discovery and disentanglement of aligned pattern clusters. Proteomes. 2018;6(1):10.
8. Wong AK, Sze-To AHY, Johanning GL. Pattern to knowledge: deep knowledge-directed machine learning for residue-residue interaction prediction. Nat Sci Rep. 2018;8(1):2045–322.
9. Zhou P-Y, Sze-To A, Wong AK. Discovery and disentanglement of aligned residue associations from aligned pattern clusters to reveal subgroup characteristics. BMC Med Genet. 2018;11(5):103.
10. Zhou P-Y, Wong AK, Sze-To A. Discovery and disentanglement of protein aligned pattern clusters to reveal subtle functional subgroups. In: 2017 IEEE international conference on bioinformatics and biomedicine (BIBM). Kansas City: IEEE; 2017.
11. Samek W, Wiegand T, Müller K. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models; 2017. arXiv preprint arXiv:1708.08296.
12. Voosen P. How AI detectives are cracking open the black box of deep learning. Science;2017. https://www.sciencemag.org/news/2017/07/howai-detectives-are-cracking-open-black-box-deep-learning.
13. Wong AK, Li GC. Simultaneous pattern and data clustering for pattern cluster analysis. IEEE Trans Knowl Data Eng. 2008;20(7):977–23.
14. Zhou P-Y, Li GC, Wong AK. An effective pattern pruning and summarization method retaining high quality patterns with high area coverage in relational datasets. IEEE Access. 2016;4:7847–58.
15. Wong AK, Zhou P, Sze-To A. Discovering deep knowledge from relational data by attribute-value association. In: Proc. 13th Int. Conf. Data Min. DMIN'17; 2017.
16. Cheng J, Ke Y, Ng W. δ-Tolerance closed frequent itemsets. In: Sixth international conference on data mining (ICDM'06), Hong Kong;2006, p. 139–48. https://doi.org/10.1109/ICDM.2006.1. https://ieeexplore.ieee.org/abstract/document/4053042?casa_token=wN7NYMxevd8AAAAA:0w6-FStj5rjV-QHj7ncpXGvBj4wylQ-hkDFjL_vKq_YywE1KFlCeGdEsOXj0u_uXbASEL2s.
17. Li J, Liu G, Wong L. Mining statistically important equivalence classes and delta-discriminative emerging patterns. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining; 2007, p. 430–9. https://dl.acm.org/doi/abs/10.1145/1281192.1281240?casa_token=gzcpJh2miJEAAAAA%3Abh-XHMSL35m8CR8CThhu8qR0MH5A5lr2xfGAGR2FGFXSKtNgBogO0qAB6T7ozLEw4-Y5kL1goZs.
18. Wong AK, Wang Y. Pattern discovery: a data driven approach to decision support. IEEE Trans Syst Man Cybern Part C Appl Rev. 2003;33(1):114–24.
19. Abdelhamid N, Thabtah F. Associative classification approaches: review and comparison. J Inf Knowl Manag. 2014;13(03):1450027.
20. U. M. L. Repository. Thoracic surgery data data set, 13 November 2013. Available: http://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data.
21. Agrawal R, Tomasz I, Arun S. Mining association rules between sets of items in large databases. ACM SIGMOD Rec. 1993;22(2):207–16.
22. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018;2(10):719–31.
23. Liang HY, Tsui B, Xia H, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med. 2019;25:433–8.
24. Ali L, Zhu C, Golilarz NA, Javeed A, Zhou M, Liu Y. Reliable Parkinson's disease detection by analyzing handwritten drawings: construction of an unbiased cascaded learning system based on feature selection and adaptive boosting model. IEEE Access. 2019;7:116480–9.
25. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learning Technol. 2011;2(1):37–63. https://www.researchgate.net/publication/276412348_Evaluation_From_precision_recall_and_Fmeasure_to_ROC_informedness_markedness_correlation.
26. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):6.
27. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. In: 2010 20th international conference on pattern recognition; 2010.
28. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
29. Branco P, Torgo L, Ribeiro R. A survey of predictive modelling under imbalanced distributions; 2015. arXiv preprint arXiv:1505.01658.
30. Ling CX, Li C. Data mining for direct marketing: problems and solutions. In: Kdd; 1998.
31. He H, Ma Y. Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons; 2013. https://books.google.ca/books?hl=zh-TW&lr=&id=CVHx-Gp9jzUC&oi=fnd&pg=PT9&dq=Imbalanced+learning:+foundations,+algorithms,+and+applications&ots=2iKpHklq5m&sig=Zr0x96yUy_-HOJrEmqEL25k3fXk#v=onepage&q=Imbalanced%20learning%3A%20foundations%2C%20algorithms%2C%20and%20applications&f=false.
32. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. Learning from imbalanced data sets. Berlin: Springer; 2018. p. 1–377.

33. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res. 2017;18(1):559–63.
34. Napierala K, Stefanowski J. Types of minority class examples and their influence on learning classifiers from imbalanced data. J Intell Inf Syst. 2016;46(3):563–97.
35. Zhuang DE, Li GC, Wong AK. Discovery of temporal associations in multivariate time series. IEEE Trans Knowl Data Eng. 2014;26(12):2969–82.
36. Wang S. Mining textural features from financial reports for corporate bankruptcy risk assessment. M.Sc. Thesis, Systems Design Engineering, University of Waterloo, Waterloo; 2017.

Publisher's Note