Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explanation and Prediction of Clinical Data with Imbalanced Class Distribution based on Pattern Discovery and Disentanglement

Peiyuan Zhou (choupeiyuan.ca@gmail.com)
University of Waterloo https://orcid.org/0000-0001-6651-0079
Andrew K.C. Wong
University of Waterloo

Research article

Keywords: Pattern Discovery, Disentanglement, Aligned Residue Associations, Aligned Pattern Clusters, Subgroup Characteristics

Posted Date: May 13th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-28409/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published on January 9th, 2021. See the published version at https://doi.org/10.1186/s12911-020-01356-y.

Abstract

Background Statistical data analysis, especially the advanced machine learning (ML) methods, have attracted considerable interest and application in clinical practices. First, the interpretability of the diagnostic/prognostic results will bring confidence to doctors, patients and their relatives in therapeutics and clinical practice. Furthermore, from the clinical aspect, when the datasets are imbalanced in diagnostic categories, the ordinary ML methods might produce results overwhelmed by the majority classes diminishing prediction accuracy. Hence, it is desirable to have a method that could produce explicit transparent and interpretable results in decision-making, even for data with imbalanced groups.

Methods In order to interpret the clinical patterns and conduct diagnostic prediction of patients, we present our new 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), 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 and succinct set of statistically significant high-order patterns from clinical data for interpreting and predicting the disease class of the patients even for small and rare groups.

Results Experiments on synthetic and thoracic clinical dataset showed that cPDD can 1) discover fewer 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 discover all patterns implanted in the data, display 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 explainable method in solving the imbalanced class problem shows its great potential to clinical data analysis for years to come.

Background

Clinical diagnostic decisions have a direct impact on the outcomes and treatment of patients in the clinical setting. As large volumes of biomedical data are being collected and becoming available for analysis, there is an increasing interest and need in applying machine learning (ML) methods to diagnose diseases, predict patient outcomes and propose therapeutic treatments. For example, Deep Learning (DL) has been successful in assisting analysis and classifying medical scans, X-rays, etc. Although DL is generally considered as a black box [1] lacking transparency to interpret why a decision is made, yet for these forms of visual data, users with cognition ability are able to relate the targets to the input 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. If the patterns inherent in the relational data, though not visualized, are succinctly related to the targets, existing ensemble algorithm, 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 [2]. Hence, existing ML approaches on relational data are still encountering difficult problems concerning transparency, low data volume, and/or imbalance classes [3] [4].

To render transparency and interpretability, Decision Tree, Frequent Pattern Mining or Pattern Discovery were proposed. For decades, *Frequent Pattern Mining*[5] [6] [7] is an essential data mining task to discover knowledge in the form of association rules from relational data [7]. However, as revealed in our recent work [8] [9] [10], the Attribute Value Association (AVA) forming patterns of different classes/targets could be entangled due to multiple entwining functional characteristics inherent in the source environments. Hence, the patterns discovered directly from the acquired data may have overlapping or functionally entwined AVAs as observed from our recent works [8] [10].

Hence, in this paper, we present a new classification method based on Pattern Discovery and Disentanglement (PDD) with the capability to tackle this problem. We particularly focus on imbalanced class problem since it is still challenging most of the traditional ML methods.

The cPDD algorithm is briefly described in Fig. 1. From a clinical relational dataset **R** say with *N* attributes, the frequency of occurrences for all distinct Attribute-Value (AV) pairs (or second order Attribute-Value Associations (AVAs)) are first obtained. Then, the frequency of occurrences is turned into a statistical measure known as adjusted statistical residual (SR) [7] which accounts for the deviation of that frequency from the default model if the AVs in the AVA pair is statistically independent. So then, a matrix of SRs is obtained, and each SR represents the statistical interdependency of an AV pair. In this matrix, each row represents a AV-vector with its coordinates representing the SR values of the AVA it associates with other AV's corresponding to the VS of the column vector. This matrix is thus referred to as the AVA Statistical Residual Vector Space (SRV). The next step is applying principal component decomposition (PCD) to decompose the SRV into different principle components (PCs) and re-project the projections of the AV-vectors on each PC after the transformation 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. We refer a PC with its RSRV together as a Disentangled Space (DS). Since the number of DSs is as large as the number of AVs, cPDD uses a DS-Screening Algorithm to select a small set of DSs denoted by **DS*** = $\{DS_i^*\}$ if the maximum SR in the RSRV of that DS exceeds a set statistical threshold (e.g. 1.96 in 95% confidence interval). The AVs with statistically significant AVAs will form Attribute-Value Clusters (AV-Clusters) in the PCs reflecting groups of strongly associating AVs.

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 a small number of AV-Clusters from a small set **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 in AV clusters[10][11] and subgroups in 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, cPDD significantly reduces the variance problem and relates more specific patterns to the targets. Unlike traditional 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 uncorrelated AVA source environment will enhance prediction and classification, particularly effective for data with imbalanced classes.

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 input (in terms of inherent patterns) and output (decision targets/classes) relations are not that obvious, particularly when the correlation of signs, symptoms, test results of the patients could be the manifestation of multiple factors[3] [12]. Hence, this poses a challenge to DL in clinical application. Another concern is on the transparency and the assured accuracy[3] [12]. As for transparency, DL is generally considered as a black box [1]. Although ML methods like ensemble algorithm, 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 [2]. Hence, to render transparency and interpretability, Decision Tree, Frequent Pattern Mining or Pattern Discovery were proposed. Since rules discovered by Decision Tree is guided by class labels, it is unlikely to discover associations between attributes when class labels are not available. Furthermore, as revealed in our recent work [8] [9] [10], associations 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 [7], support, confidence or statistical residuals[6] [7], may have overlapping or functionally entwined AVA patterns captured in the 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 robust or comprehensive.

cPDD that we proposed in this paper has solved the fundamental pattern entanglement problem. It is proposed to meet the clinical challenges posed above. It intends to provide clinical results explainable to clinicians using a small number of patterns discovered from the disentangled sources 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 Spaces orthogonal to those of the majority classes.

Novelty and Contributions

cPDD extends our recent work [10] on AVA disentanglement to the discovery of statistically significant high-order patterns in AVA disentangled spaces. It provides robust and succinct interpretation and achieves from clinical data with anomalies and imbalance class distribution more specific and precise prediction. 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 imbalanced.

Methods

In this section, we extend our previous work, Attribute-Value Association Discovery and Disentanglement Model (AVADD) [10] [11] [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.

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	Principle Component Decomposition
RSRV	Re-projected SRV
DS	Disentangled Space
DS*, DS*	Selected Disentangled Space, the selected set

First, we denote the input data as **R**, which contains *N* attributes, denoted as A = $\{A_1, A_2, \ldots A_N\}$. For a numerical value, say A_n, we partition its values into I_n bins using Equal Frequency algorithm [16] to transform numerical AVs into discrete values. After transforming, the input data can be denoted as *N*

attributes, and each attribute (A_n) is denoted as $A_n = \{A_n^1, A_n^2, \ldots A_n^{I_n}\}$. Then, cPDD is implemented in the following five steps.

1.

Statistical Data Analysis: The first step is the same as in high-order pattern discovery[7] which uses statistical method to construct an Adjusted Statistical Residual Vectors (SRV) to represent the statistical weights of all the AVA pairs obtained from **R.** Thus, each item of SRV is denoted as a SR ($\mathrm{A_n^i} \leftrightarrow \mathrm{A_{n'}^j}$), which represents the adjusted residual between two AVs ($\mathrm{A_n^i} \leftrightarrow \mathrm{A_{n'}^j}$). SR of an AVA pairs defined as (*SR(* $A_n^i \leftrightarrow A_{n'}^j$*,)* = SR_{ij} = $\frac{o_{ij}-e_{ij}}{\sqrt{e_{ij}}}$) to account for the deviation of its observed frequency of occurrences against the expected frequency of occurrences if the AVs in the pair are statistically independent. Thence, SRV is an $I \times I$ matrix of SRs, where $I = \sum_{n=1}^{N} I_n$ represents the total number of distinct AVs. Generally, the significant associations can be selected according to the threshold obtained from the hypothesis test of statistically significant SR. For example, when its SR > 1.96, the association can be treated as positively significant with a 95% confidence level.

2.

Acquisition of AVA Disentangled Spaces: For AVA disentanglement, Principal Component Decomposition (PCD) is applied to discompose the SRV into *k* PCs, denoted as PC=$\{PC_1, PC_2, \ldots PC_k \ldots PC_N\}$, where PC_k is a set of projections of the AV vectors from the SRV, where PC_k= { $PC_k\left(A_n^i\right) | n = 1,2, \ldots N, i = 1, \ldots I_n\}$. We then re-project the projections of the AV-vectors captured in the PCs to a new SRV with the same basis vectors and call it a Reprojected SRV (RSRV). We then refer all the PCs and the their corresponding RSRVs = $\{RSRV_1, RSRV_2, \ldots RSRV_k, \ldots RSRV_N\}$ as the AVA disentangled spaces (DSs) where $RSRV_k$ is the re-projected result on PC_k via $RSRV_k = SRV \bullet PC_k \bullet {PC_k}^T$. 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 vector in the RSRV. The coordinates of these AV vectors in the RSRV represent the SRs of the AVAs captured in the PCs. We refer a PC with its RSRV as a Disentangled Space (DS). Figure 2 shows a DS (PC and RSRV) for 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 1.96 at confidence level of 95%). 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 a AV clusters is that each statistically significant AVA Pair must have one of its AVs having a significant AVA with other AV in the cluster. In the RSRV (Fig. 2), the green shade shows an AVA pair associating with other AVs.

4.

Pattern Discovery: High-order patterns can be discovered through identifying pattern candidates from an AV cluster growing process. Formally, we denote a high-order pattern as $P_j$which consists of a subset of AVs with size$\geq$2. We use the adjusted residual[3] 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[17][18] in the pattern discovery process. There might me 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, 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[19] [20] 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 data and the clinical dataset with imbalanced classes. In this section, we present the experimental results and exemplify the capability of cPDD in the analysis.

Materials

Dataset 1: Synthetic Dataset

We generated stochastically a 2100 × 10 matrix with first column as the class label and others as attributes with character values from a uniform distribution. This represents a random relational dataset with attributes independence to each other. We then embedded patterns of three different classes „ and 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 C1 and C2; A3H and A4M are entangled in C1 and C3; A5B and A6J are entangled in C2 and C3. 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 is implemented as one with imbalanced class distribution with 1000 entities pertaining to C2 and C3 each, and 100 entities pertaining to C1.

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

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 [21]. 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 Risk1Y. Risk1Y = T if the patient died. In this dataset, the class distribution is imbalanced with 70 cases being Risk1Y = T and 400 cases being Risk1Y = F.

Analysis I – Discovery and Display of Explicit Patterns for Explanation

In Analysis I, we applied cPDD on both Synthetic and Thoracic Datasets. First, we compared the discovered patterns obtained in cPDD, Apriori[22] (a typical frequent pattern mining method) and a high-order pattern discovery method for discrete-value data (HOPD)[7] which was our early work closely resembling the PD reported in [11] [10]. Figure 4 and Fig. 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.

From the pattern discovery result of cPDD on the synthetic data, we observe that:

i.
A small set of AV-Clusters (Fig. 5A) was discovered. From these clusters a comprehensive set of patterns (Fig. 5B), each of which associates with a different class or subgroup in a DS*, was discovered.
ii.
The AV-clusters (Fig. 5A), also representing the unions of all comprehensive patterns discovered within subgroups in a DS* (Fig. 5B), can be considered as the summarized patterns consisting of statistically significant AVA subsets for a high-level interpretation with details given in the comprehensive patterns.

iii.
cPDD discovered and displayed both the summarized patterns (Fig. 4A) and the detailed patterns (Fig. 4B) associating with both classes (Fig. 6A) whereas HOPD could not due to the overwhelming number of overlapping and entangled patterns disclosed; whereas Apriori, after fine-tuning the level of support and confidence *(sup = 10%, con = 10%)*, discovered the patterns associated with the rare class (Fig. 6A).
Furthermore, when comparing the implanted patterns with the pattern discovery result using cPDD (Fig. 4C), cPDD reveals all patterns with correct classes in disentangled spaces except one in P2 as it has (A2D, A3G, A5B, A6J). Although the pattern is same with the implanted patterns, but it shares with sub-pattern in P3 showing 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 discovered a large number of patterns where most of them are redundant and overlapping, while some of them are associating with class labels, but others are with the noise columns A7, A8 and A9.

From the pattern discovery result on the Thoracic dataset, we observed similar phenomena as described in item (i) to (iii) as in the case on the synthetic data if we replace Fig. 4 and Fig. 6A with Fig. 5 and Fig. 6B respectively. Here we would like to highlight some interesting observations in the Thoracic case.

Figure 5A gives four AV-Clusters, two in each AVA disentangled Space (DS1 and DS2). Each AV-cluster contains the union of all the patterns discovered in different subgroups (Fig. 5B). They are explainable. It is interesting to observe that cPDD discovered only 9 patterns (Fig. 4B and Fig. 6B) each of which pertains to a distinct implanted class within distinct AV clusters in different DS* except those in AV Cluster 2 of DS2 encompassing patterns of both C2 and C3. However, their statistical strength SR was relatively low. This demonstrates that cPDD is able to display precise statistical/functional details of the pattern implanted in the data –– an astounding evidence of its explanation efficacy.

Figure 6B shows the comparison results for the Thoracic data. First, we should note that when the number of patterns is large with considerable redundant and overlapping patterns, it is difficult to interpret the pattern outcomes relevant to problem. The number of patterns obtained by Apriori and HOPD are both large. Apriori outputs the patterns from dataset only if the class labels are given. HOPD can output all the patterns discovered among the growing set of the candidate patterns, 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. Next, we try to examine whether the Apriori and HOPD are able to discover the patterns associated with the minority class. For Apriori, the result depends 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.

In summary, this experimental result shows that cPDD is able to discover fewer patterns with specific association to the classes/source-environment in support of easy/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 posts a novel approach of Explainable AI (XAI) [23] [24] quite different from the existing model-based 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. For the imbalanced class problem, usually the targeted group is the minority group. 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 accuracy criterion. It should be evaluated by the Precision and Recall of the minority class and the F1-Score which summarizes the harmonic mean of both the majority and the minority groups. Thus, F1-score = 0 if the number of true positive TP = 0. In this experiment, the average Precision, Recall and F1-Score obtained from the 20 10-fold cross validation of the three classification methods are obtained and shown in Fig. 7. The comparison results showed that cPDD outperformed the other two classification methods.

When comparing with a recent pattern discovery method PD, we used the result on the same dataset from the work reported in [2]. We noted that cPDD outperformed PD in both precision and F-measure. The PD method[2] acquired lower precision rate than that of cPDD, but a F1-Score of 0.3±0.01 which is close to that obtained by cPDD (F1-Score = 0.31±0.02). We also noticed that Decision Tree misclassified all the test cases since it did not discover any rule for the cases with Risk1Y = T.

Conclusion

As a pattern discovery model on imbalanced data, the experimental results on synthetic and Thoracic data have shown that cPDD renders superior prediction performance and explainability since it produces and uses much smaller set of succinct disentangled patterns. All the results it obtains are statistically robust, comprehensive, displayable in succinct concise and precise representation for experts' interpretation. It also overcomes the limitations of lack of transparency [12] as well as the problem of imbalanced class[3][12][4] [25]. As a clinical data analysis tool on relational data, it has a significant advantage over 'blackbox' ML algorithms as the outputs of cPDD is both explainable and transparent, the two major challenges of interpretability and applicability[23] confronting ML today. The experimental result on synthetic data and clinical data with high imbalanced class ratios shows that cPDD does have a better interpretability and prediction performance for minority target. 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[23] [24].

In our future work, cPDD will be developed to apply to unstructured data (e.g. text and sequences) [8] [26] by extracting AVAs directly from them as shown in our early work [27]. Moreover, for performance

improvement, parallel computing strategy will be introduced to handle bigger data and further speed up the computational time.

List Of 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	Principle Component Decomposition
RSRV	Re-projected SRV
DS	Disentangled Space
DS*, DS*	Selected Disentangled Space, the selected set

Declarations

Funding

Publication costs were funded by NSERC Discovery Grant (xxxxx 50503-10275 500)

Availability of data and materials

The datasets in this study are available from the corresponding author on reasonable request.

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.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Voosen, "How AI detectives are cracking open the black box of deep learning," *Science,* 2017.
2. Chan, Y. Li, C. Chiau, J. Zhu, J. Jiang and Y. Huo, " Imbalanced target prediction with pattern discovery on clinical data repositories. ," *BMC medical informatics and decision making,,* vol. 17, no. 1, p. 47, 2017.
3. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature medicine,* vol. 25, no. 1, pp. 44-56, 2019.
4. Aggarwal and S. Sathe, "Bias Reduction in Outlier Ensembles: The Guessing Game," in *Outlier Ensembles*, Springer, 2017.
5. Naulaerts, W. Bittremieux, T. Vu, W. Vanden Berghe, B. Goethals and K. Laukens, "A Primer to frequent itemset mining for bioinformatics," *Briefings in bioinformatics,* vol. 16, no. 2, pp. 216-231, 2015.
6. C. Aggarwal and J. Han, Frequent pattern mining, Springer, 2014.
7. K. Wong and Y. Wang, "High-Order Pattern Discovery from Discrete-Valued Data," *IEEE Transaction On Knowledge System,* vol. 9, no. 6, pp. 877-893, 1997.
8. P.-Y. Zhou, A. E. Lee, A. Sze-To and A. K. Wong, "Revealing Subtle Functional Subgroups in Class A Scavenger Receptors by Pattern Discovery and Disentanglement of Aligned Pattern Clusters," *Proteomes,* vol. 6, no. 1, p. 10, 2018.
9. K. Wong, A. H. Y. Sze-To and G. L. Johanning, "Pattern to Knowledge: Deep Knowledge-Directed Machine Learning for Residue-Residue Interaction Prediction," *Nature Scientific Reports,* vol. 8, no. 1, pp. 2045-2322, 2018.
10. P.-Y. Zhou, A. Sze-To and A. K. Wong, "Discovery and disentanglement of aligned residue associations from aligned pattern clusters to reveal subgroup characteristics," *BMC medical genomics,* vol. 11, no. 5, p. 103, 2018.
11. P.-Y. Zhou, A. K. Wong and A. Sze-To, "Discovery and Disentanglement of Protein Aligned Pattern Clusters to Reveal Subtle Functional Subgroups.," in *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on. IEEE*, Kansas City, MO, USA, 2017.
12. Samek, T. Wiegand and K. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296,* 2017.

13. K. Wong and G. C. Li, "Simultaneous pattern and data clustering for pattern cluster analysis," *IEEE Transactions on Knowledge and Data Engineering,* vol. 20, no. 7, pp. 977-923, 2008.
14. P.-Y. Zhou, G. C. Li and A. K. Wong, "An Effective Pattern Pruning and Summarization Method Retaining High Quality Patterns With High Area Coverage in Relational Datasets," *IEEE Access,* vol. 4, pp. 7847-7858, 2016.
15. A. K. Wong, P. Zhou and A. Sze-To, "Discovering Deep Knowledge from Relational Data by Attribute-Value Association," in *Proc. 13th Int. Conf. Data Min. DMIN'17.*, 2017.
16. A. K. Wong and D. C. Wang, "DECA: A discrete-valued data clustering algorithm.," *IEEE transactions on pattern analysis and machine intelligence,* no. 4, pp. 342-349, 1979.
17. J. Cheng, Y. Ke and W. Ng, "\delta-Tolerance Closed Frequent Itemsets," in *Data Mining, 2006. ICDM'06. Sixth International Conference on. IEEE*, 2006.
18. J. LI, G. Liu and L. Wong, "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. ACM*, 2007.
19. A. K. Wong and Y. Wang, "Pattern discovery: a data driven approach to decision support.," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* vol. 33, no. 1, pp. 114-124, 2003.
20. N. Abdelhamid and F. Thabtah, "Associative classification approaches: review and comparison," *Journal of Information & Knowledge Management,* vol. 13, no. 03, p. 1450027, 2014.
21. U. M. L. Repository, "Thoracic Surgery Data Data Set," November 2013. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data.
22. R. Agrawal, I. Tomasz and S. Arun, "Mining association rules between sets of items in large databases," *Acm sigmod record,* vol. 22, no. 2, pp. 207-216, 1993.
23. K.-H. Yu, A. L. Beam and I. S. Kohane, "Artificial intelligence in healthcare," *Nature biomedical engineering,* vol. 2, no. 10, pp. 719-731, 2018.
24. H. Y. Liang, B. Tsui, H. Xia and etc., "Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence," *Nature Medicine,* vol. 25, pp. 433-438, 2019.
25. K. Napierala and J. Stefanowski, "Types of minority class examples and their influence on learning classifiers from imbalanced data," *Journal of Intelligent Information Systems,* vol. 46, no. 3, pp. 563-597, 2016.
26. D. E. Zhuang, G. C. Li and A. K. Wong, "Discovery of temporal associations in multivariate time series," *IEEE Transactions on Knowledge and Data Engineering,* vol. 26, no. 12, pp. 2969-2982, 2014.
27. S. Wang, "Mining Textural Features from Financial Reports for Corporate Bankruptcy Risk Assessment," M. Sc. Thesis, Systems Design Engineering, University of Waterloo, Waterloo, 2017.

Figures

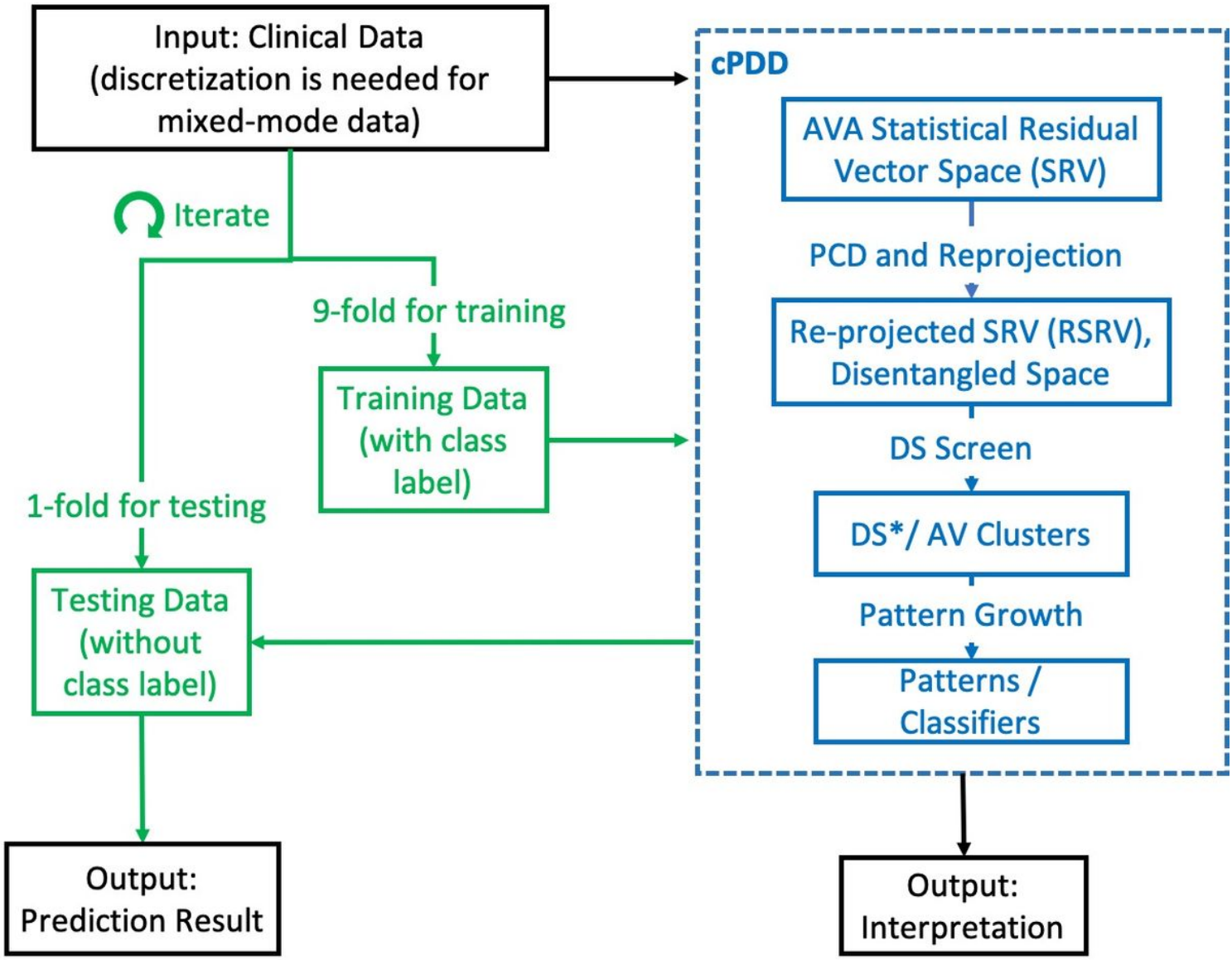
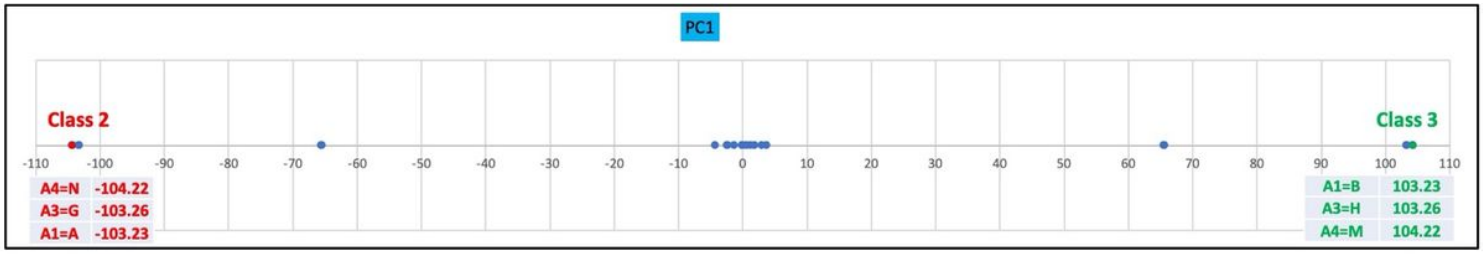


Figure 1

Overview of cPDD for Interpretation and Prediction



RSRV1	Class=C3	A4=M	A3=H	A1=B	A6=F	A2=D	A8=O	A9=O	A7=Q	A9=P	A7=P	Class=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

Figure 2

An illustration of DS* with two AV clusters in the first PC using synthetic data. As displayed in the Re-projected SRV (RSRV), class 2 and class 3 are disentangled in the first PC and the first RSRV, and they can be grouped clearly. In the figure, the order of the AVs in the RSRV is reversed to correspond to those in the PC plot.

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)**

Figure 3

Attribute Description of Thoracic Dataset

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			

(A)

		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			

(B)

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

(C)

Figure 4

cPDD Pattern Discovery Result from Synthetic Dataset. (A) In the First and the Second DS* two AV-clusters were discovered. (B) Detailed Patterns associating with different classes were discovered from the above two AV-Clusters. (C) Comparison between implanted patterns and cPDD's output.

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	

(A)

		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	

(B

Figure 5

Pattern Discovery Result of Thoracic Dataset using cPDD (A) In both First and the Second Disentangled Space, two AV Clusters corresponding to Risk1=T and RISK1=F were discovered (B) Detailed Patterns discovered from the above tow AV Clusters.

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

(A)

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

(B)

Figure 6

(A) Comparison of Pattern Discovery Result on Synthetic Dataset (B) Comparison of Pattern Discovery Result on Thoracic Dataset

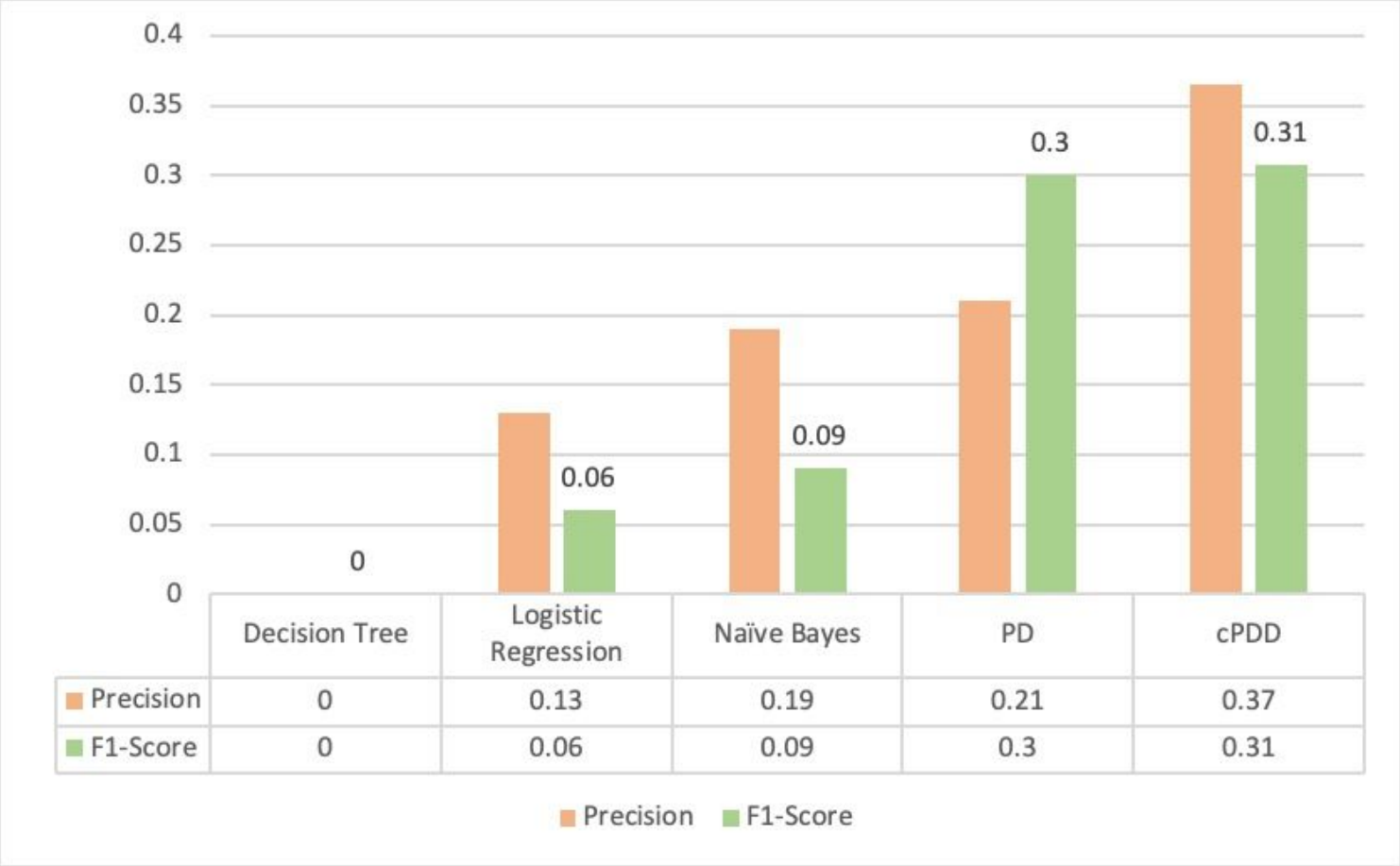


Figure 7

Average Classification Result from 20-times 10-fold cross validation on Thoracic Dataset.