Pattern Discovery and Disentanglement for Clinical Data Analysis

by

Peiyuan Zhou

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
System Design Engineering

Waterloo, Ontario, Canada, 2020

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In recent years, machine learning approaches have important empirical successes on analysing data such as images, signals, texts and speeches with applications in biomedical and clinical areas. However, from the perspective of modelling, many machine learning methods still encounter crucial problems such as the lack of transparency and interpretability. Frequent Pattern Mining or Association Mining methods intend to solve the problem of interpretability, but they also encounter serious problems such as requiring exhaustive search and producing overwhelming numbers of patterns. From the perspective of data analysis, they do not render high prediction accuracy particularly for data with low volume, rare or imbalanced groups, rare cases or biases due to subtle overlapping or entanglement of the statistical and functional associations at the data source level. Hence, Professor Andrew K.C. Wong and I have developed a novel Pattern Discovery and Disentanglement (PDD) Method to discover explicit patterns and unveil knowledge from relational datasets even encompassing imbalanced groups, biases and anomalies. The statistically significant high-order patterns, pattern clusters and rare patterns are discovered in the disentangled Attribute Value Association (AVA) Spaces. They may be embedded in a relational dataset but overlapping or entangled with each other so that they are masked or obscured at the data level. The patterns discovered from the disentangled association source can be used for explicitly interpreting the original data, predicting the functional groups/classes and detecting anomalies and/or outliers. When class labels are not given, pattern/entity clustering can be more effectively discovered from the disentangled attribute value association (AVA) space than from the original records. The objective of this Master Thesis is to develop and validate the efficacy of PDD for genomic and clinical data analysis using a) protein sequence data, b) public clinical records from UCI dataset and c) a clinical dataset obtained from the School of Public Health and Health Systems at the University of Waterloo. The experimental results with superior performance in unsupervised and supervised learning than existing methods are presented in interpretable knowledge representation frameworks, interlinking the AVA disentangled sources, patterns, pattern/entity clusters and individual entities. In the clinical cases, it reveals the symptomatic patterns of individual patients, disease complexes/groups and subtle etiological sources. Hence it will have impacts in machine learning on genomic and clinical data with broad applications.

Acknowledgements

Firstly, I would like to express my sincere appreciation to my MASc. supervisor, *Prof. Andrew K.C. Wong,* for his gentle and patient supervision. His generous encouragement and support give me confidence to overcome difficulties during my MASc. studies. I am deeply grateful for his insightful comments, carful revision for my research paper and this thesis, which inspire me to devote myself fully to the research work. I would say that I sincerely consider myself very fortunate to learn from him to become a critical thinker and an innovative problem solver.

I would like to thank Prof. Zahid A. Butt, Prof. Helen Chen (School of Public Health and Health Systems, The University of Waterloo), and Dr. Puiwing Wong for giving me many valuable suggestions and revision on my research paper revision. And also, I also would like to thank all the professors for teaching my lectures. Their lectures and tutorials inspire me for my research work. I also would like to thank the university to provide vibrant environment for my study. I also would like to thank George and Rena in School of Public Health and Health Systems, and all my colleagues and friends in the System Design Engineering for their friendship and encouragement.

Finally, I would like to thank my parents, their love supports me throughout the time of my student life. I also would like to thank my husband, Jieming Li and my son, Sixuan Li. Without their love and support, I would not complete this thesis. So, thanks go to my family, to whom I piously dedicate this thesis.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Background and Research Motivations

Clinical diagnostic decisions have a direct impact on the treatment and the outcomes 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 from the data.

Machine Learning and Deep Learning have important empirical successes on analysis of data such as images, signals, texts and speeches with outcomes akin to human cognition and discernment. Although they are 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 applying to relational datasets (**R**) for comprehensive data analysis such as clinical analysis and practice, the interpretability of these methods is still a challenge [1] [2]. For example, the characteristics associations (or Attribute-Value Associations AVA) may overlap with many "either-or" cases, further complicating the decision and interpretation in ML. That is to say, if the patterns inherent in the relational data, though not visualized, are succinctly related to the targets, existing ensemble algorithms, such as Boosted SVM, or Random Forest could produce good predictive results, but not if the AVAs in the different targets are entangled. Moreover, the underlying patterns in support of the decision are still opaque and uninterpretable by clinicians [3].

In addition, as noted in [2], AI today still focuses on improving accuracy, but provides little interpretation. It may lead to overdiagnosis in the healthy population, increasing the burden handled by health care systems instead of relieving it [3]. Current Explainable AI studies focus on model explanation, but not the interpretation of the model in clinical uses. The latter is highly desired in the clinical context [3].

Hence, the challenges in applying machine learning techniques difficult ML problems still being encountered in clinical practice are listed as follows:

a) lacking transparency for understanding the throughputs and outputs [2] [4];

b) difficulty in identifying anomalies [2] [5]; and

c) getting biased results when the data size is small or the class distribution is severely imbalanced as in the case of the rare diseases [6] [7].

These impasses have to be overcome before AI could solve some crucial data analytic problems in the medical area.

1.2 Problem and Solution

To address model's transparency and interpretability, Decision Tree, Frequent Pattern Mining and Pattern Discovery were proposed. For decades, *Frequent Pattern Mining* [8] [9] [10] is an essential data mining technique to discover knowledge in the form of association rules from relational data [10]. However, they usually produce an overwhelming number of overlapping/redundant patterns/rules coming from entwined classes/groups [11]. These patterns/rules are hard to partition/summarize [11] [12] [13] to reveal precise "knowledge" inherent in the source environment, making interpretation difficult and lowering prediction accuracy. In addition, as shown in our recent work [14] [15] [16], the Attribute Value Association (AVA) forming patterns of different classes/targets could be entangled due to multiple entwining functional characteristics, i.e. class labels, 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 [14] [16]. Furthermore, existing ML approaches on relational data are still encountering difficult problems concerning transparency, low data volume, and/or imbalanced classes [2] [5].

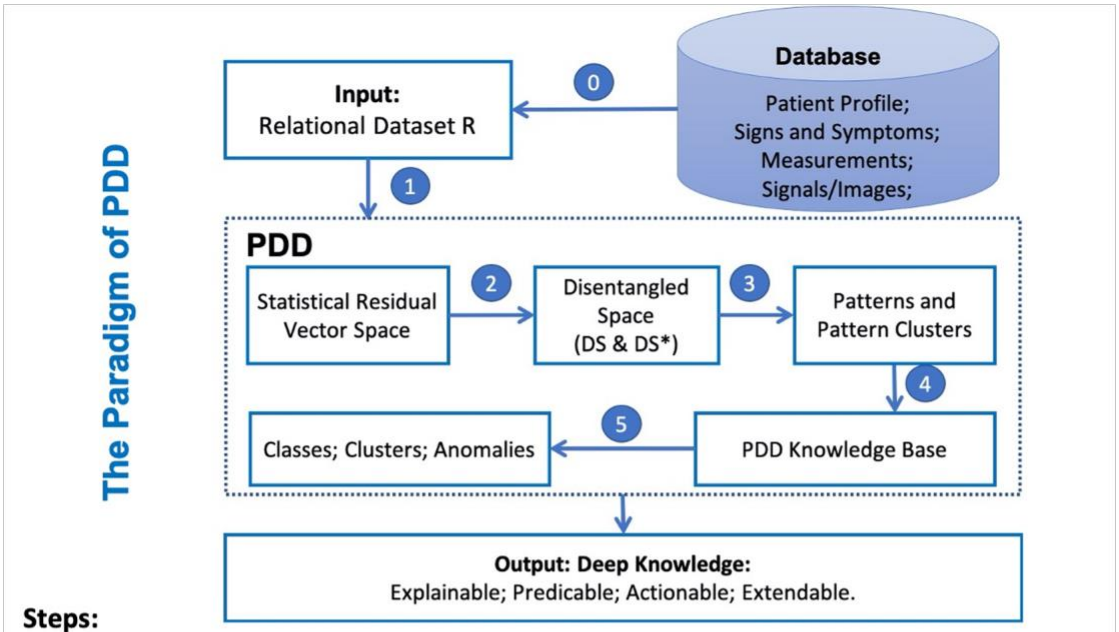
Therefore, the objectives of this study are:

1) Developing a pattern discovery approach with explainability and applying it in clinical practice;
2) Predicting the class information for the clinical data using patients' profile and symptomatic characteristic patterns;
3) Clustering the patients according to their characteristic patterns even when class label is not given;
4) Detecting the anomaly cases (e.g. the rare cases or outliers) from the clinical data using the above pattern discovery process.

Table 1 Terminology

Terminology	Brief Definition	Medical Examples
Pattern Entanglement	AVAs forming patterns could pertain to different classes. They could be co-occurring/overlapping among entities and are hard to separate for prediction and interpretation. They are entangled.	Signs, symptoms, test results and patient's physical profile from multiple diseases or etiological causes; mixed indicators from treatment/drug responses.
Disentanglement	A process to project AVAs pertaining to different classes/origins onto orthogonal disentangled spaces DS*, from which succinct patterns could form different pattern clusters and entity clusters.	Patterns are related to more specific pathological and etiological causes; and, rare cases or anomalies could be traced back to their entangled origins which could be related to certain disease classes/causes.
Deep Knowledge	Obscured knowledge interlinking DS*, patterns and entities. They are not visualized/recognizable at the data level.	The subtle causes of a disease; manifestation of multiple disorders; misdiagnoses/mis-prognoses, best treatments identified.
PDD-Knowledge Base (PDD-KB)	A unified knowledge representation consisting a Summary-KB and a Comprehensive-KB interlinking 3 parts: Disentangled Space (DS*), Patterns and Entities – to support ML and interpretation.	DS*: disease causes, syndromes, disorders, cyberchondria; etc. Pattern: signs-symptoms groups, patient's profiles, best treatments. Entity: patients' records PDDKB can link them together.
EID-Intersection of an AVA	The set of entities each containing that AVA. It is equivalent to frequency count of the AVA in R.	Patients sharing the same group of indicators.
Anomaly and outlier, entities	Anomalies: patterns beyond present knowledge; Outliers: entities contain no discovered patterns at certain statistical threshold but could reveal rare patterns/clues at deeper levels. Rare Cases: entity found in classes not as labeled.	Anomalies: Patients found with new conditions not previously identified. Outliers: Patient with no identified conditions of a disease complex. Rare Cases: Patients misdiagnosed or with misinformation in the records.

Figure 1 Overview of PDD Applying to Clinical Data



Steps:

0) Obtain characteristic features from various forms of medical data; input them in a relational dataset **R** for comprehensive pattern discovery and analysis.

1) From **R,** PDD obtains an AVA Frequency Matrix (AAVAFM), and converts it into a Statistical Residual Vector Space (SRV). The statistical residual (SR) of an AVA is a measure of the deviation of its observed frequency from the base case where the AVs are mutually independent.

2) PDD then obtains the disentangled AVA spaces (DSs), each consisting of a Principal Component (PC) and its Re-projection Statistical Residual Vector Space (RSRV). Amongst all DSs, PDD selects those, denoted by DS*, with SR(s) in RSRV exceeding a prescribed statistical threshold.

3) PDD uses AV clustering for PD and entity clustering and discovers patterns from the AV clusters by a pattern statistic test, which greatly reduces pattern search complexity.

4) PDD then produces a unified knowledge representation: PDD Knowledge Base (PDDKB), interlinking patterns, entities and source environments up to individual pattern and entity level.

5) PDD accomplishes difficult ML tasks such as grouping patterns/entities without relying on class information; classifying/clustering entities from **R** even with imbalanced classes/groups and presence of noise/biases/anomalies; improving classification rate by removing bias and errors and displaying analytical results for interpretation and further knowledge exploration/organization.

Hence, the data-driven exploratory method Pattern Discovery and Disentanglement (PDD) [17] developed by Andrew K.C. Wong and me is adopted in this thesis to discover robust/succinct patterns with statistical analysis and implicit functional clues to explain the underlying relation and augment scientific exploration. PDD discovers deep knowledge from relational datasets. In PDD, Deep knowledge refers to the subtle function/relations/associations that are masked/inconspicuous at the data level due to source entanglements, but can be discovered and represented in a compact and succinct knowledgebase which displays a much smaller set of explicit patterns linking to individual entities as well as the classes or underlying causes that begat those specific association patterns. PDD provides a new scientific perspective as it overcomes several crucial limitations of the current AI methods plagued by biases, imbalanced classes, rare groups, anomalies and lack of transparency. Such problems raise concerns in medical/clinical ML applications. Hence, the explainability addressed in PDD attempts to meet the clinical challenges rather than pose just a technical discourse. It intends to provide clinical results explainable to a clinical practitioner, understood by the patients with statistical symptomatic characteristic patterns in support of diagnostic and detection rare cases from imbalanced clinical data and so on.

Figure 1 gives an overview of PDD applying to a clinical data setting and Table 1 provides terminology descriptions with medical examples.

1.3 Organization of Thesis

The rest of this thesis is organized as follows.

In chapter 2, the thesis presents a summary of existing work on machine learning, especially pattern discovery models, for clinical data analytics. The literature survey on clinical data analysis is introduced in the same chapter. The advantages and drawbacks of major existing works are described.

Chapter 3 describes the details of each step that PDD takes to tackle pattern discovery problems. From the displayable explicit patterns discovered, PDD's result can also be used for interpretation, clustering, classification of disease complexes/patients and detection of rare cases.

From Chapter 4 to Chapter 6, the results of interpretation, unsupervised learning, rare cases detection, and supervised learning are precisely presented. For evaluating the performance of PDD, its interpretation capability, clustering result and prediction results are compared with those obtained by

its existing counterparts in Chapter 4, Chapter 5, and Chapter 6 respectively. The experimental results demonstrate that the proposed algorithm is effective not only with respect to the theoretical construct and algorithmic robustness and efficacy, but also to its practical applicability. Especially, for some clinical dataset taken directly from hospitals/clinics with imbalanced classes, PDD always outperforms other traditional machine learning algorithms.

Finally, chapter 7 draws a conclusion from the clinical application addressing the notion of whether or not, and to what extent, the proposed method can be used for clinical practices. It also presents the evaluation of the current limitations of PDD in this study and suggests directions for future research.

Chapter 2
Literature Survey

2.1 Machine Learning on Clinical Data Analysis

Due to the ever-expanding digital data sources, it is an obvious trend of using artificial intelligence and machine learning algorithms on clinical data analysis [18]. For example, in the area of drug discovery, some machine learning methods are used to predict pharmaceutical properties of molecular compounds [19] [20]. In addition, in order to enable faster diagnoses and tracking of disease progression, pattern recognition and segmentation techniques are applied to medical images, such as retinal scans, pathology slides and body surfaces [21] [22]. Furthermore, in order to extend the applications using new predictive models, deep learning techniques are applied to the combined genomic and clinical data [23] [24]. However, the central problem of such models is that they are regarded as black-box models. And even if the underlying mathematical principles of such models are understood, they lack an explicit declarative knowledge representation. Hence, they have difficulty in generating the underlying explanatory structures [25]. Translating machine learning models to clinical practices needs establishing clinicians' trust [26], which requires transparency in the algorithms and the presented results. As for transparency, DL is generally considered as a black box [1]. Although ML methods like ensemble algorithm, such as Boosted SVM (BSI) for imbalanced data, or Random Forest are good at prediction, their classification results are highly opaque and difficult for the clinicians to interpret [3].

2.2 Pattern Discovery Models

Considering transparency and interpretability, pattern mining and discovery models can provide the explicit detailed patterns discovered from the datasets [3]. Decision Trees and Forests, Frequent Pattern Mining or Pattern Discovery were proposed. For decades, *Frequent Pattern Mining* [9] [8] [10] is an essential data mining task to discover knowledge in the form of association rules from relational data [10]. Most of them are based on the likelihood, the weight of evidence [10], support, confidence and/or statistical residuals [9] [10]. However, as revealed in our recent work, [14] [15] [16], associations discovered from relational data 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 attribute values as observed from our recent works [14] [16]. This notion is further validated and exemplified by our synthetic experiment in Chapter 4,

resulting in serious pattern overlapping/redundancy [12] [11] and uncontrollable entwinement. These overwhelming yet statistically legitimate patterns are difficult to partition and summarize [12] [11] [13]. Hence, patterns discovered by current pattern mining and discovery methods are unable to reveal and use the deep knowledge inherent in the data without pattern disentanglement.

In the past decades, many methods [9] have been developed for discovering high-order patterns or mining frequent rules [8] [10]. However, their performance is sensitive to manually set thresholds such as *support* and *confidence* and the patterns they discovered were overwhelmed [11] with overlapping/redundant patterns due to their combinatorial nature and the possible entanglement in the source environment. Optimal ways of identifying, partitioning and grouping patterns were lacking. Pattern clustering [11] attempts to group them according to their similarity for better visualization and interpretation. However, they typically produce too many pattern clusters and complex cluster configurations. Although our later work, such as pattern pruning and summarization [11] [12], attempted to produce more succinct representations when similar patterns are clustered into groups, such approaches still require intensive search and it is hard to find effective similarity measure to direct the clustering. After a long search for better optimal methods, the key problem lies in the entanglement of AVA due to multiple governing functions/factors in the source environment [15]. Most of the previous methods cannot discriminate the subtle variations inherent in the entangled sources. Such observation poses a challenge to current DL/ML models since, in these cases, the input-to-output relationship upon which they are based, if not properly disentangled, is not succinct to render good solutions. Hence, in the problem as revealed in a recent study, both the above-mentioned approaches lack effective ways to disentangle the statistics obtained from the data coming from subtle multiple entangled sources [16].

Furthermore, existing ML approaches are still encountering difficult problems concerning transparency, low data volume, imbalance classes and presence of anomalies, biases and rare samples [2] [5]. Hence, PDD was developed to meet these challenges.

2.3 Deep Learning and PDD

Today, deep learning (DL) and frequent pattern mining are two commonly used methodologies for data analysis. Undoubtedly, DL is a powerful tool for learning complex, cognitive tasks related to vision, signals, speech and text where humans can cognitively relate the input to the output. By this token,

successful analyses/classifications on medical scans, X-rays, retinal images, ECG, etc. have been reported. However, in a more general healthcare data analytics based predominantly on clinically recorded numeral and descriptive data, the input and output relations are not always obvious, particularly when the correlation of signs, symptoms, test results of a patient could be the manifestation of multiple diseases. Hence, this poses a challenge to DL in clinical application. Another concern is on transparency and assured accuracy [2] [4]. As for assured accuracy, DL usually requires a large size of data with a broad base of coverage and experts' supervision to ensure high predictive accuracy. Without such assurance, sometimes researchers found that simpler, cheaper and more useful data modelling could render better results [2]. When applying to precision medicine and discovering rare cases, DL still requires the strong support of prior knowledge and feature engineering [27], demanding enormous human effort. Furthermore, the medical community still wishes to understand how machines learn [28] and what sort/level of accountable knowledge they could discover and offer. Clinicians are looking for an AI system which is explainable and accountable, able to discover anomalies and rare cases [2] [5] without sacrificing predictive accuracy. While empirical science relies on statistics, PDD renders deeply embedded statistical patterns [12] in scientific exploration.

Chapter 3

Pattern Discovery and Disentanglement

3.1 Methodology of Pattern Discovery and Disentanglement

This chapter presents the PDD methodology applying to relational datasets. At the outset, the values of numeral attributes in **R** are quantized into interval values via entropy maximization [10]. The Overview in Figure 1 renders greater details of the algorithm. Figure 2 (A) shows how PDD accomplishes the proposed tasks in six enumerated steps, with definitions, algorithmic description and justification given for each.

Step 1 constructs an AV-Address Table (AT) by attaching to each AV in **R** a list of Entity Identities (EID) containing that AV**.**

Step 2 constructs an AVA Frequency Matrix (AVAFM) [16] by obtaining the AVA frequency of each AV-pair as the number of intersecting entities in the AV-pair instead of through searching **R** exhaustively. From the AVAFM, an AVA Statistical Residual Vector Space (SRV) is constructed for later AVA Disentanglement by converting each frequency into a statistical residual --- a statistical measure accounting the deviation of the observed frequency of the AVA from that if the AVs in it are independent.

Step 3 disentangles the entangled AVA statistics by applying Principal Component Decomposition (PCD) on the SRV [15] to obtain the Principal Components (PCs) and their corresponding Re-projected SRVs (RSRVs) with the same set of SRV basis vectors [14] [15]. A PC with its corresponding RSRV is referred as an AVA Disentangled Space (DS).

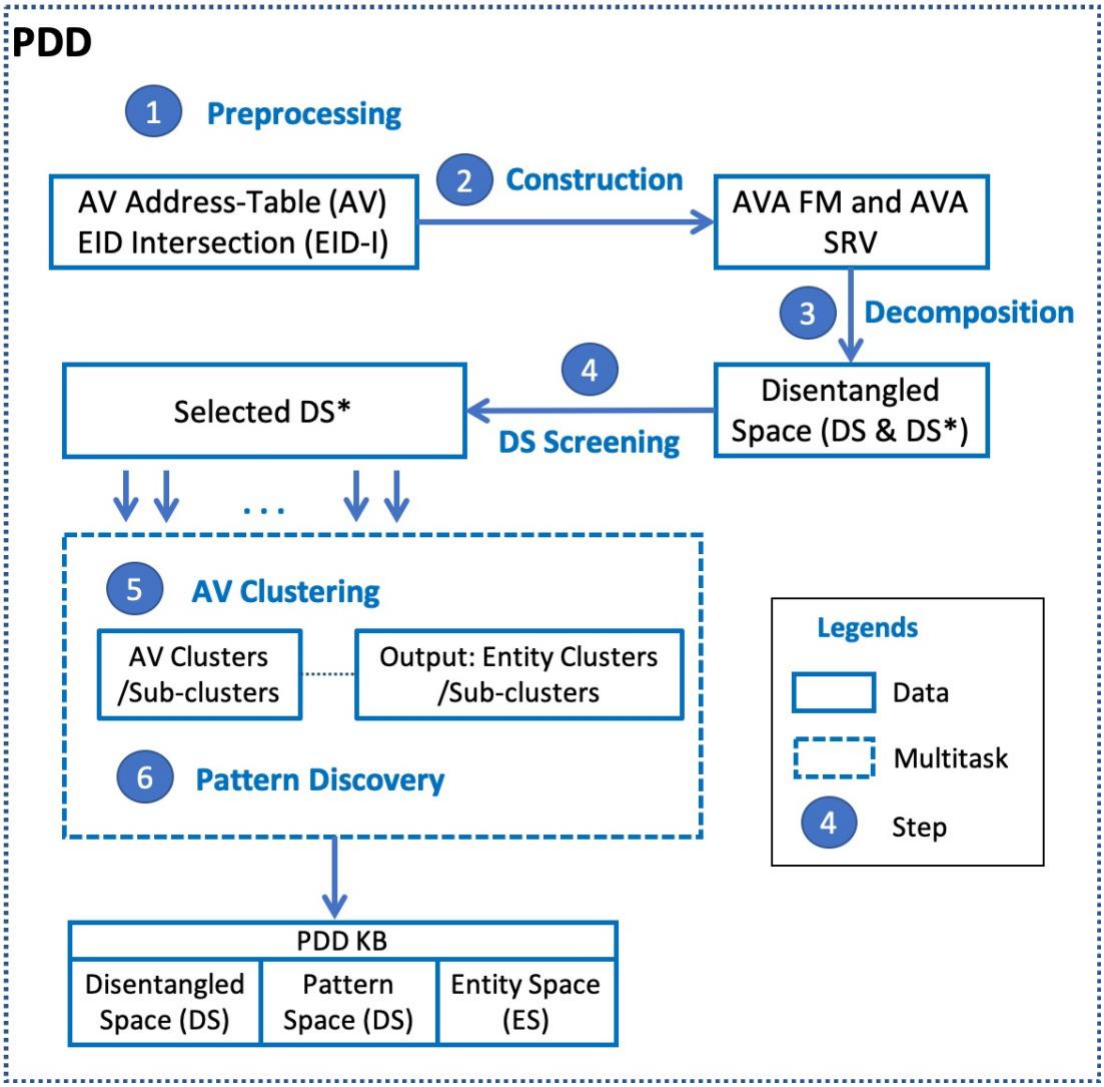
Since in PCD, the number of DS is as large as that of AVs, Step 4 selects only the statistically significant DS, denoted by DS*, if the maximum SR in its RSRV exceeds a prescribed SR threshold. In general, only a very small set of DSs is selected.

In Step 5, on each DS*, the AV-Clustering Algorithm is applied to obtain one or more AV clusters from the AV projections captured in the PC. As an example, shown in Figure 2 (B), in the PC, only two

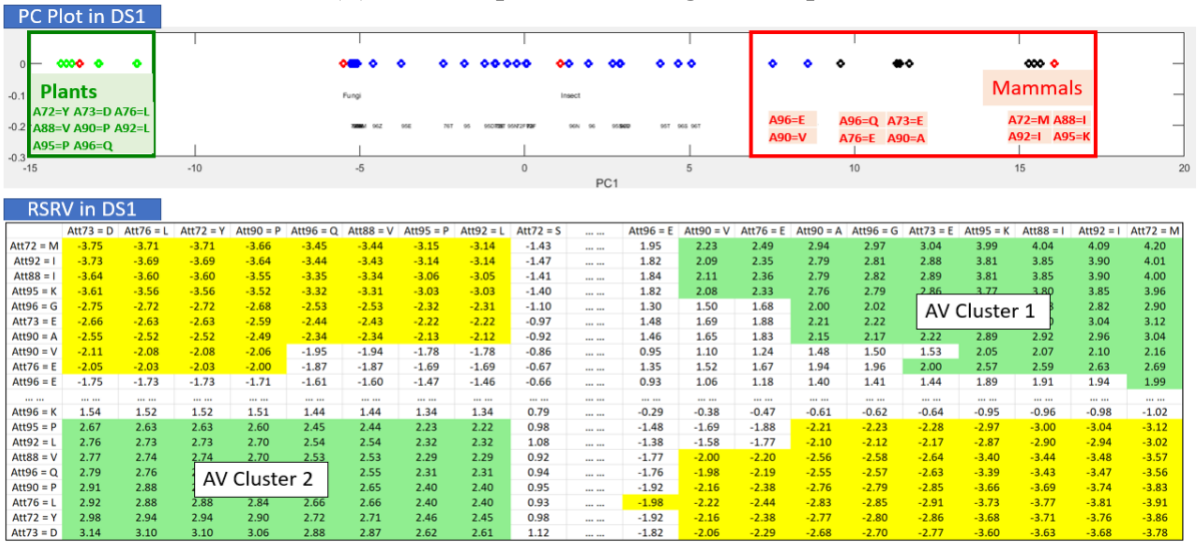
small sets of AVs associating with Plants and Mammals are obtained at the far ends, reflecting their contribution to the eigenvalue in the PC while their strong AVAs are captured.

Figure 2

(A) Schematic Overview of the PDD algorithm



(B) An Example of Disentanglement Space



	Att73 = D	Att76 = L	Att72 = Y	Att90 = P	Att96 = Q	Att88 = V	Att95 = P	Att92 = L	Att72 = S	... ...	Att96 = E	Att90 = V	Att76 = E	Att90 = A	Att96 = G	Att73 = E	Att95 = K	Att88 = I	Att92 = I	Att72 = M
Att72 = M	-3.75	-3.71	-3.71	-3.66	-3.45	-3.44	-3.15	-3.14	-1.43	... ...	1.95	2.23	2.49	2.94	2.97	3.04	3.99	4.04	4.09	4.20
Att92 = I	-3.73	-3.69	-3.69	-3.64	-3.44	-3.43	-3.14	-3.14	-1.47	... ...	1.82	2.09	2.35	2.79	2.81	2.88	3.81	3.85	3.90	4.01
Att88 = I	-3.64	-3.60	-3.60	-3.55	-3.35	-3.34	-3.06	-3.05	-1.41	... ...	1.84	2.11	2.36	2.79	2.82	2.89	3.81	3.85	3.90	4.00
Att95 = K	-3.61	-3.56	-3.56	-3.52	-3.32	-3.31	-3.03	-3.03	-1.40	... ...	1.82	2.08	2.33	2.76	2.79	2.86	3.77	3.80	3.85	3.96
Att96 = G	-2.75	-2.72	-2.72	-2.68	-2.53	-2.53	-2.32	-2.31	-1.10	... ...	1.30	1.50	1.68	2.00	2.02				2.82	2.90
Att73 = E	-2.66	-2.63	-2.63	-2.59	-2.44	-2.43	-2.22	-2.22	-0.97	... ...	1.48	1.69	1.88	2.21	2.22				3.04	3.12
Att90 = A	-2.55	-2.52	-2.52	-2.49	-2.34	-2.34	-2.13	-2.12	-0.92	... ...	1.46	1.65	1.83	2.15	2.17	2.22	2.89	2.92	2.96	3.04
Att90 = V	-2.11	-2.08	-2.08	-2.06	-1.95	-1.94	-1.78	-1.78	-0.86	... ...	0.95	1.10	1.24	1.48	1.50	1.53	2.05	2.07	2.10	2.16
Att76 = E	-2.05	-2.03	-2.03	-2.00	-1.87	-1.87	-1.69	-1.69	-0.67	... ...	1.35	1.52	1.67	1.94	1.96	2.00	2.57	2.59	2.63	2.69
Att96 = E	-1.75	-1.73	-1.73	-1.71	-1.61	-1.60	-1.47	-1.46	-0.66	... ...	0.93	1.06	1.18	1.40	1.41	1.44	1.89	1.91	1.94	1.99
... ...	... ...	... ...	... ...	... ...	... ...	... ...	... ...	... ...	... ...	... ...	... ...	... ...	... ...	... ...	... ...	... ...	... ...	... ...	... ...	... ...
Att96 = K	1.54	1.52	1.52	1.51	1.44	1.44	1.34	1.34	0.79	... ...	-0.29	-0.38	-0.47	-0.61	-0.62	-0.64	-0.95	-0.96	-0.98	-1.02
Att95 = P	2.67	2.63	2.63	2.60	2.45	2.44	2.23	2.22	0.98	... ...	-1.48	-1.69	-1.88	-2.21	-2.23	-2.28	-2.97	-3.00	-3.04	-3.12
Att92 = L	2.76	2.73	2.73	2.70	2.54	2.54	2.32	2.32	1.08	... ...	-1.38	-1.58	-1.77	-2.10	-2.12	-2.17	-2.87	-2.90	-2.94	-3.02
Att88 = V	2.77	2.74	2.74	2.70	2.53	2.53	2.29	2.29	0.92	... ...	-1.77	-2.00	-2.20	-2.56	-2.58	-2.64	-3.40	-3.44	-3.48	-3.57
Att96 = Q	2.79	2.76	2.76	2.72	2.55	2.55	2.31	2.31	0.94	... ...	-1.76	-1.98	-2.19	-2.55	-2.57	-2.63	-3.39	-3.43	-3.47	-3.56
Att90 = P	2.91	2.88	2.88	2.84	2.65	2.65	2.40	2.40	0.95	... ...	-1.92	-2.16	-2.38	-2.76	-2.79	-2.85	-3.66	-3.69	-3.74	-3.83
Att76 = L	2.92	2.88	2.88	2.84	2.66	2.66	2.40	2.40	0.93	... ...	-1.98	-2.22	-2.44	-2.83	-2.85	-2.91	-3.73	-3.77	-3.81	-3.91
Att72 = Y	2.98	2.94	2.94	2.90	2.72	2.71	2.46	2.45	0.98	... ...	-1.92	-2.16	-2.38	-2.77	-2.80	-2.86	-3.68	-3.71	-3.76	-3.86
Att73 = D	3.14	3.10	3.10	3.06	2.88	2.87	2.62	2.61	1.12	... ...	-1.82	-2.06	-2.29	-2.68	-2.70	-2.77	-3.60	-3.63	-3.68	-3.78

All other AVs only with weak AVAs with other AVs are close to the center of the PC (with weak AVA and thus irrelevant in that DS*). The small set of strong AVAs are shown in the RSRV in green blocks.

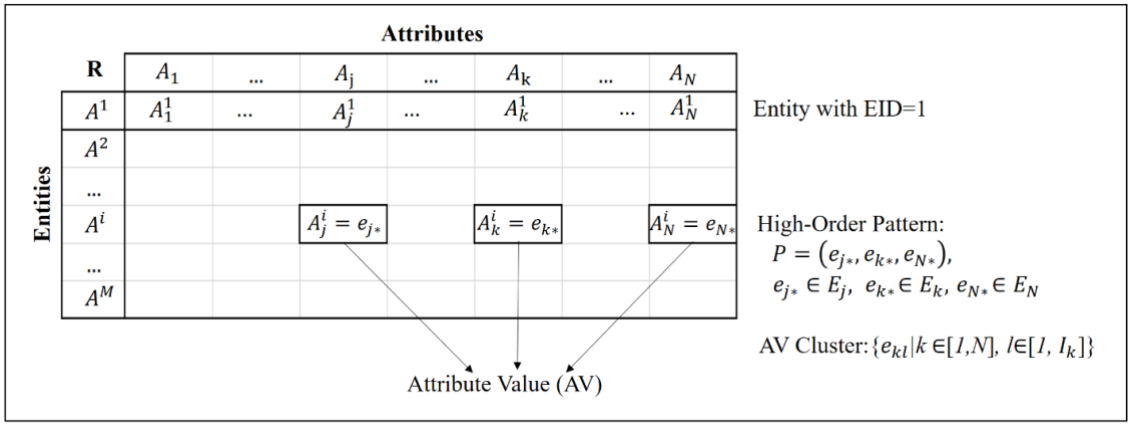
In Step 6, a pattern discovery algorithm is used to discover high-order statistically significant patterns from the AV Clusters in each DS* instead of through extensive search from **R**. When an AV-cluster/sub-cluster passes the pattern statistical test, it is referred to as a pattern. With the closeness of AVs in the 1-Dimensional PC space and the easily tracked SR value of the AVAs in 2-Dimensional RSRV, the process of AV-Cluster identification and pattern confirmation is fast and effective in comparison with the laborious sorting and counting from the entire dataset in traditional High-Order Pattern Discovery (HOPD) [11] [12] [10]. The patterns discovered from an AV cluster/sub-cluster form a pattern group (PG)/pattern-subgroup (SubPG) respectively. The entities covered by an AV cluster form an entity group (EG).

Finally, the knowledge obtained from Step 6 is combined into a unified representation framework referred to as the PPD Knowledge Base (PDDKB). PDDKB interlinks DS*, patterns, and all entities in **R**, in both a summarized form and also a form with comprehensive details, i.e. the complete list of patterns discovered by PDD. From the PDDKB, all the deep knowledge can be extracted and organized for the pattern discovery posterior tasks such as class prediction, entity clustering, etc. without relying on class information.

3.1.1 Input and Preprocessing

The input relational dataset is denoted as **R,** which is an $N \times M$ relational dataset, where N is the number of attributes, $A_1, \ldots, A_N$ and M the number of entities A_i, $i = 1,\ldots, M$ each of which is assigned with a unique Entity ID (EID) i. As described in Figure 3, **R** contains N attributes, and an attribute can take on either a numerical or a categorical value, so **R** could be a mixed-mode relational dataset with M N-tuples of mixed-mode data (numerical / categorical). For a mixed-mode relational dataset, the numerical values are needed to be discretized into quantized intervals. Equal Frequency method [29] is usually used.

Figure 3 An illustration of relational dataset R, EID, AV Cluster and Pattern



After discretization, by treating each interval as a categorical value, **R** contains categorical values only. Let e_{ji} represent the *ith* value of the *jth* attribute in **R**. Let $E_j = \{e_{j1}, e_{j2}, \ldots e_{jI_j}\}$ be the set of all possible categorical (discrete) attribute values (AVs) of A_j, with I_j being the total number of possible values of the j_{th} attribute. Therefore, the total number of AVs across all attributes is $\sum_{j=1}^{N} I_j$. the ith entity is denoted as A^i, and the AV of the jth attribute of the ith entity as A_j^i and $A_j^i \epsilon\, E_j$. Figure 3 summarizes the notations used in this section.

To reduce the computational complexity for Steps 2 to 6, the Entity Address Table (**AT**) is created. In the AT, for each distinct AV in **R**, it lists the Entity ID (EID) of the entities in **R** containing that AV.

Definition 1. The Entity Address Table AT. **AT** is a table with $T = \sum_{j=1}^{N} I_j$ slots corresponding to all possible distinct AVs, i.e. $\bigcup_{j=1}^{N} E_j$, in **R**. The slot associated with an AV e_{kl} contains the list of EIDs L_{kl} of the entities containing e_{kl}, where $L_{kl} = \{i = 1, \dots, M | A_k^i = e_{kl}\}$.

3.1.2 Construct SRV

In Step 2, a statistical method is used to construct an Attribute-Value Association Frequency Matrix (AVAFM) to represent the frequency of occurrences of all the AV pairs obtained from **R.**

Definition 2. AVA Relative Frequency Matrix AVAFM. AVAFM is a $T \times T$ matrix of AVA relative frequencies between two AVs, say e_{ni} and $e_{n'j}$. The frequency entry of the matrix is $f_{ni \leftrightarrow n'j} = \frac{|L_{ni} \cap L_{n'j}|}{M}$.

Through **AT**, AVAFM can be constructed using the cardinality of the EID Intersection (EDI-I) of the AV pairs, i.e., $| L_{ni} \ L_{n'j} |$ of the AVA pair, say e_{ni} and $e_{n'j}$, instead of exhaustively searching and counting the AVA pairs directly from **R.**

Then the Adjusted Statistical Residual (SR), denoted by $SR_{ni \leftrightarrow n'j}$ between an AV pair $e_{ni \leftrightarrow n'j}$, is used to measure whether an AVA frequency (say between e_{ni} and $e_{n'j}$) in the FM is statistically significant or not from Eqn (1).

$$SR_{ni \leftrightarrow n'j} = \frac{r_{ni \leftrightarrow n'j}}{\sqrt{v_{ni \leftrightarrow n'j}}} \quad (1)$$

where $r_{ni \leftrightarrow n'j}$ represents the standardized residual of $e_{ni \leftrightarrow n'j}$;

$r_{ni \leftrightarrow n'j} = \frac{Occ(e_{ni \leftrightarrow n'j}) - Exp(e_{ni \leftrightarrow n'j})}{\sqrt{Exp(e_{ni \leftrightarrow n'j})}}$;

$v_{ni \leftrightarrow n'j}$ represents the maximum likelihood estimate of the variance of $r_{ni \leftrightarrow n'j}$ and

$v_{ni \leftrightarrow n'j} = \mathrm{Var}(r_{ni \leftrightarrow n'j}) = \ (1 - \frac{|L_{ni}|}{M} * \frac{|L_{n'j}|}{M})$;

$Occ(e_{ni \leftrightarrow n'j}) = f_{ni \leftrightarrow n'j} * M$ (total number of occurrences for $A_{nk} = e_{ni}$ and $A_{n'l} = e_{n'j}$)

$Exp(e_{ni \leftrightarrow n'j}) = \frac{|L_{ni}| * |L_{n'j}|}{M}$; (expected frequency) and M is the total number of entities.

Thus, the Space SRV, a $T \times T$ matrix with the entry of $SR_{ni \leftrightarrow n'j}$ for each AV pair $e_{ni \leftrightarrow n'j}$, is constructed as that reported in [16].

3.1.3 **Obtain Disentangled Space (DS)**

A greater detailed exposition on how to obtain DS can be found in our previous work [15]. Firstly, the row vector associating with the attribute value "*a"* is denoted as the *a*-vector. To disentangle the SRV, the Principal Component Decomposition (PCD) is applied on the SRV to obtain a set of PCs. The projections of the *a*-vectors on each PC are then re-projected onto a new SRV referred to as its corresponding RSRV (using the same basis vectors). After PCD, a set of k PCs denoted as PC = $\{PC_1, PC_2, \ldots PC_k\}$ is obtained where PC_k is a set of projections of the *a*-vectors in a one-dimensional space from SRV.

Then, as the coordinates of a row AV-vector in the SRV represents the statistical weights of that AV associating with another AV denoted by its column vector, PCD maps the *a*-vectors onto different uncorrelated (orthogonal) PC axes. Then if the projections of the *a*-vectors are mapped onto a new SRV with the same set of basis vectors, they are the corresponding *a*-vectors in the RSRV revealing the AVAs captured by that PC in the SRV but now re-projected onto the RSRV.

Hence, the relation (AVAs) of the *a*-vectors captured by that PC will then be reflected in the corresponding RSRV. This is the essence of the SRV disentanglement. With the same basis vectors for all RSRVs as those of the SRV, a unified representation framework with the same set of basis vectors as SRV can be used to interpret the AVAs captured in different PCs. By adopting a consistent notation, the subscript k in $RSRV_k$ corresponds to that in PC_k.

3.1.4 **Obtain Significant Second Order AVAs.**

Since the number of DSs (PCs or RSRVs) is as large as that of AVAs and each PC is independent to others, to obtain only the statistically significant second-order AVAs from each DS, a DS Screening Algorithm is devised to select only DS (denoted by DS*) that contains statistically significant SRs in their RSRV for PD. The selection of DS* is a great reduction of space complexity for PD. It is much succinct, robust and meaningful than choosing DSs based on variance (eigenvalue) of the PCs. Algorithm 1 presents the pseudocode for DS Screening.

Algorithm 1: DS Screening

Input: All RSRV= $\{RSRV_1, RSRV_2, \ldots RSRV_n\}$, each $RSRV_k = \{RSR_{ni \leftrightarrow n'j}\}$; *sig (=1.96)*
Output: Selected DS = $\{DS_1^*, DS_2^*, \ldots\}$

%For each $RSRV_k$, the task can be completed in parallel multitasking setting

BEGIN
For each $RSRV_k$ in RSRV
  For each $RSR_{ni \leftrightarrow n'j}$ in $RSRV_k$
    If $RSR_{ni \leftrightarrow n'j} > sig$
      Add e_{ni} and $e_{n'j}$ in DS_k^*
    End
  End
  Add DS_k^* in Selected DS
Return Selected DS.
END

3.1.5 **AV Clustering**

Once a small set of DS*s is selected, AV clusters and sub-clusters are obtained for pattern discovery and entity clustering in an independent and parallel multitasking setting.

> ***Definition 3. Attribute-Value (AV) Clustering.*** A process finds one or two disjoint clusters in each DS_k^* through a linear search of AV subsets RSRV such that each subset contains AVs within must have a significant AVA with other AVs in it.

Thus, AV cluster(s) are incrementally grown from PC and RSRV. Figure 2(B) exemplifies the AV-Clustering process. The AVs are grouped in the same cluster in the PC because each of them has at least one statistically significant AVA (from RSRV in green shade) with another AV within the cluster. The pseudo code of AV Clustering in DS* is presented in Algorithm 2.

To introduce AV sub-clusters in a hierarchical manner, AVs in each AV clusters could be agglomerated by a similarity matrix identified through the degree of the overlapping AVs between entities. It uses similarity measures between all AVs pairs, say e_{ni} *and* $e_{n'j}$, in one DS* denoted as *sim(*$e_{ni}, e_{n'j}$*)=* $|cov(e_{ni}) \cap cov(e_{n'j})|$*, where cov(*$e_{ni}$) and *cov(*$e_{n'j}$) represent the entities covered by e_{ni} and $e_{n'j}$ respectively. With the cardinality of EID-Intersection of their coverage as a distance measure, a complete-linkage hierarchical clustering algorithm is applied to obtain the clusters when a threshold which represents the lower boundary of similarity between two clusters is met. Then, to examine the

deep knowledge discovered from the data (the data space), it is would like to see how entities are grouped based on the AVAs obtained in the disentangled space DS*. Like AV cluster, an entity cluster is composed of the entities containing correlated AVs in an AVA disentangled space while being grouped into AV clusters.

Algorithm 2: AV Clustering

Input: $DS_k^* = \{e_{kl} | k\epsilon[1,N], l\epsilon[1,I_k]\}$, $RSRV_k$, op
Output: $AVCluster = \{AVCluster_1, \ldots AVCluster_n\}$

% The selected significant AVAs are ranked by SR values
Initial setting: add two AVs of the first AVA in $AVCluster_1$
For each significant AVA ($e_{ni \leftrightarrow n'j}$) in $RSRV_k$
  For each existed $AVCluster_k$
    If e_{ni} (or $e_{n'j}$) has been in $AVCluster_k$
      add $e_{n'j}$ (or e_{ni}) in $AVCluster_k$
    End
  End
  If e_{ni} *and* $e_{n'j}$ are not in any existed $AVCluster_k$
    create $AVCluster_{k+1}$
    add e_{ni} *and* $e_{n'j}$ to $AVCluster_{k+1}$
  End
End
Return $AVCluster = \{AVCluster_1, \ldots AVCluster_n\}$

3.1.6 High-Order Pattern Discovery

After obtaining AV clusters and sub-clusters in different DS*s, a pattern discovery procedure applying statistical pattern hypothesis test to each AV cluster/sub-cluster is developed to discover high-order patterns. Figure 3 gives an example of three AVs co-occurring on the same entities. If their frequency of co-occurrences on same entities in **R** exceeds the statistical threshold, they are considered as a statistically significant pattern (Figure 3**Error! Reference source not found.**) based on the theoretical notion proposed in [10]**.**

Definition 5. High-Order Pattern. A high-order pattern P_j consists of a subset of AVs with size ≥ 2, such that the frequency of their co-occurrences on the same entities in **R** deviates significantly from the random default model, i.e. that the distribution of AVs is equi-probable and they are independent in their occurrences.

Figure 3 shows an example of a high-order pattern. In PD, *Adjusted Residual* for a candidate pattern derived from the frequency of co-occurrences of a high order AVA is used in the hypothesis test for determining whether the AVA is a statistically significant pattern. The SR for a pattern P_j is derived as

$$SR(P_j) = \frac{R(P_j)}{\sqrt{V(P_j)}} \quad (2)$$

where $R(P_j) = \frac{Occ(P_j) - Exp(P_j)}{\sqrt{Exp(P_j)}}$ is the standard residual or Pearson Residuals, $Occ(P_j) = |\bigcap_{e_{nj} \in E_n} L_{nj}|$ is the frequency of its co-occurrences on the same entities; and $Exp(P_j) = \mathrm{M} \prod_{e_{nj} \in E_n} \frac{Occ(j)}{M}$ is the expected occurrences on the same entities, and $\mathrm{V}(P_j) = 1 - \prod_{e_{nj} \in E_n} \frac{Occ(P_j)}{M}$ is the standard deviation of all the residuals, M is the number of tuples of **R.**

In order to keep the discovered patterns non-redundant, only delta-closed patterns [13] [30] are accepted in the pattern discovery process. Delta-Closed Patterns represents closed patterns with delta-tolerance. For example, for a discovered pattern P_i with the occurrence K_{P_j}, if no super-pattern P_j with the occurance $K_{P_j} \geq \delta \cdot K_{P_i}$ (δ is the *tolerance* factor) can be discovered, then P_i is called delta-closed pattern. The processes reported in [12] are adopted, which also provides the definition of sub-patterns, super-patterns and delta-closed patterns, for pattern pruning.

In summary, pattern discovery in DS* is a process by growing second-order patterns $P_i = \{e_{ni}, e_{n'j}\}$ identified in the RSRV of DS * into third, fourth and incrementally higher order patterns from the AV clusters using the adjusted residual, Eqn. (2), for pattern test. There might me more than one pattern identified in the AV cluster/sub-cluster. The discovered patterns within an AV cluster/sub-cluster will constitute a pattern cluster (PG)/pattern sub-cluster (SubPG) respectively.

3.1.7 **Output: PDD Knowledge Base**

The patterns discovered by PDD in the AV clusters of a DS* naturally become pattern clusters; and the entity clusters generated from the AV clusters will relate the entities directly within the orthogonal source environment captured in the DS*. Hence, the output is an all-in-one representation referred to

as a unified PDD Knowledgebase (PDDKB) which links the source environment, patterns and entities altogether for the ease of executing various posteriori data analytic tasks and interpretation.

If the class labels are given, the discovered patterns that are associated with class could be used for supervised learning using associative classification algorithms [31]. The classification accuracy is assessed by taking the class labels as the ground truth. If the class labels are unavailable or not given in **R**, both the pattern clusters and the entity clusters are the unsupervised outcomes of PDD. Their accuracy can be assessed after assigning the class labels back to the corresponding entities. As shown in our previous works [17] [16] [32] [14], most results from cases with or without class labels given are almost identical. For example, Figure 4 shows the discovered patterns from Heart Disease Dataset with (Figure 4(A)) or without (Figure Figure 54 (B)) class labels given in **R** are almost identical, except that one pattern, with cpt=[3 4] in SubPG2 (Figure 4(A)) of PG1 in DS1 is not in Figure 4Figure 5 (B) for the case with no class label given since a single event cpt=[3,4] alone is not a pattern.

Figure 4 Discovered Patterns for Heart Disease Dataset

(A) with class label given in **R**

DS1																
PG1	**Residual**	**Occ.**	**age**	**sex**	**cpt**	**rbp**	**sc**	**fbs**	**rer**	**mhra**	**eia**	**oldpeak**	**spess**	**nmvc**	**thal**	**class**
SubPG1	24.58	19	[29 51]							[162 202]	0	[0 0.1]	1	[0 1]	3	**1**
	21.87	8	[29 51]		[1 3]					[162 202]	0	[0 0.1]	1	[0 1]	3	**1**
SubPG2	4.88	62			[3 4]											**1**
PG2	**Residual**	**Occ.**	**age**	**sex**	**cpt**	**rbp**	**sc**	**fbs**	**rer**	**mhra**	**eia**	**oldpeak**	**spess**	**nmvc**	**thal**	**class**
SubPG1	12.05	38			[4 4]					[71 143]		[1.4 6.2]				**2**
	14.52	15	[59 77]							[71 143]		[1.4 6.2]	2	[1 3]		**2**
	12.67	16	[59 77]		[4 4]							[1.4 6.2]	2	[1 3]		**2**
	32.06	13			[4 4]					[71 143]	1	[1.4 6.2]	2	[1 3]	7	**2**
	24.88	6	[59 77]		[4 4]					[71 143]	1	[1.4 6.2]	2	[1 3]	7	**2**

(B) class label not given in **R.**

DS1															
PG1	**Residual**	**Occ.**	**age**	**sex**	**cpt**	**rbp**	**sc**	**fbs**	**rer**	**mhra**	**eia**	**oldpeak**	**spess**	**nmvc**	**thal**
SUbPG1	17.88	19	[29 51]							[162 202]	0	[0 0.1]	1	[0 1]	3
	16.1	8	[29 51]		[1 3]					[162 202]	0	[0 0.1]	1	[0 1]	3
PG2	**Residual**	**Occ.**	**age**	**sex**	**cpt**	**rbp**	**sc**	**fbs**	**rer**	**mhra**	**eia**	**oldpeak**	**spess**	**nmvc**	**thal**
SubPG1	7.06	16	[59 77]		[4 4]								2	[1 3]	7
	8.44	17	[59 77]		[4 4]					[71 143]			2	[1 3]	
	7.35	14	[59 77]		[4 4]					[71 143]		[1.4 6.2]	2		
	9.24	16	[59 77]		[4 4]					[71 143]		[1.4 6.2]		[1 3]	
	21.04	13			[4 4]					[71 143]	1	[1.4 6.2]	2	[1 3]	7
	16.41	6	[59 77]		[4 4]					[71 143]	1	[1.4 6.2]	2	[1 3]	7

The result presented in Figure 4 gives very strong support that the patterns discovered in the DS*s by PDD are much more distinctly associating with the disentangled sources reflecting the inherent characteristics of different classes/groups. Thus, without relying on prior knowledge, PDD can interlink AV clusters (function association), patterns, and entities together in a PDD Knowledge Base (PDDKB) to enhance supervised (with class labels), unsupervised (without class labels) and semi-supervised ML.

Since PDD works with/without class labels [14] [16], it is able to address ML in a more general setting, including discovering rare events/patterns in DS*s and solving the imbalanced class, bias and noisy problems. Due to the transparency of its process, contents of its throughput and output, PDD can reveal established knowledge and/or new findings inherent in the throughput/output data assisted by experts and/or supporting evidences and experimental verification.

3.2 Parameter Setting

To grow high-order patterns in a DS* from their RSRVs, two parameters are needed to be specified in PDD: 1) statistical significance threshold *sig*; and 2) delta tolerance factor $\boldsymbol{\delta}$ for pruning patterns; *sig* $SR(P_j)$ is used to assess the pattern P_j. In statistic, setting *sig* to be 1.96, corresponding to the p-value of 0.025 [33] is a common practice. The parameter δ is the sufficient fraction for a pattern to be considered as being mostly covered by its super-pattern [13] [12]. The choosing of $\delta = 0.8$ is a good practice. The *overlap* is an optional parameter used to evaluate the density of pattern sub-clusters with [0,1]. The upper bound, 1, correspond to the patterns in the same cluster cover the same entities, while the lower bound, 0, corresponding to no sub-cluster are constructed. For the *overlap*=*n* (0<n<1) corresponds to the overlapping percentage of the entities covered by the patterns in the same SubPG, not larger than $n \times 100\%$. The overlap is set to 0.5 as a reasonable choice.

3.3 Time Complexity Analysis

The time-complexity analysis of PDD, our other recent work [15] [16] and traditional High-Order Pattern Discovery (HOPD) [10] [12] for their major procedures is given as below.

1) SRV Construction: when AT is used for SRV construction, for an N x M matrix **R**, the time-complexity of PDD is O(MN) compared with O($(MN)^2$,of our previous method [10].

2) Decomposition: SRV is an $n \times n$ matrix where n is the number of AVs. The time complexities of decomposition and reprojection using PCD are O(n^3) [34] and O(n^2) respectively. Hence, the total time complexity of obtaining DS is O(n^3)+O(n^2). However, it can be reduced to O(n^2) [35] when a distributed implementation of stochastic PCD is applied.

3) DS* Screening: the complexity of searching each SR is O(n^2) if there are n attribute values.

4) Pattern discovery on each DS*: The complexity of pattern discovery is exponential [36]. However, PDD reduces the number of candidates dramatically from all entities in **R** to a very small significant space DS*, so the time complexity is also reduced to O(2^c), where c is the number of candidate AVs in DS*, which is very small in comparison with n.

In addition, the study [12] [11] shows great complexity challenges in the pattern post-analysis -- from fundamental concepts and algorithmic approaches, to existing data mining discipline. First, the computation complexity of pattern discovery is exponential O(2^N), and that of the pattern pruning is O(p^2) , where p depends on the number of patterns. p may be huge since the mined patterns are always redundant [12]. Even after pruning the mined patterns into a small number p', the complexity of pattern clustering algorithm and K-means is O(p'^3), where p>>n and p'>>n. Therefore, instead of discovering, clustering and summarizing patterns with high complexity by traditional approaches [12] [11], PDD can obtain disentangled high-order patterns, PGs, SubPGs, ECs in an all-in-one step with low complexity. Since DS* is independent of each other, the PDD computational process can be executed in a parallel multitasking setting, further improving the speed of the entire process.

Chapter 4

Interpretability and Bias Avoidance Capability of PDD

To tackle the problem of lack of transparency existing in the current machine learning approach for prediction, the experimental result in this chapter exemplifies PDD's interpretability and the bias avoidance capability of pattern discovery. In section 4.1, a synthetic data set with implanted patterns is presented and the PDD result proves that the implanted patterns are discovered and can be easily used for pattern association interpretation. Section 4.2 presents the pattern discovery results on imbalanced datasets --- one on imbalanced synthetic dataset with implanted patterns and another on imbalanced clinical datasets. The results show that the patterns are still discovered even for the minority class due to the disentanglement capability of PDD.

4.1 Pattern Discovery Result on Synthetic Dataset

The experiment on a synthetic dataset is provided to demonstrate and validate the succinct and precise association interpretation capability of the PDD. The generated synthetic dataset is a 3000 x 16 matrix with first column as the class label and others as attributes with character values stochastically generated from a uniform distribution. Then the patterns generated for three different classes C_1, C_2, and C_3 are embedded for the first ten attributes. To introduce noise and uncertainty, randomly generated character values are embedded in the other five attributes. To reveal attribute values and their association, a simple notation of AV is introduced. The notation A1A, as an example, is used to represent that the attribute A1 takes on the character value A, and A3E/F to represent that the AV A3 takes on value E or F and so forth. Some of the correlated character values on the same entities are generated from the embedded patterns for the first ten attributes. For the last five attributes, characters from {"O", "P", "Q"} are randomly embedded. The patterns implanted in **R** are given in Table 2. Note that A1A, A2C, A3E/F are entangled (overlapping) for C1 and C2; A4H, A5M, A6A/B are entangled in C1 and C3; A7I, A8J, A9G/K are entangled in C2 and C3.

Table 2 Embedded Patterns for the Synthetic Dataset

Classes	Attribute Values are Significant Associated with Class Label
C1	A1A, A2C, A3E/F, A4H, A5M, A6A/B, A7D, A8F, A9G, A10N/L
C2	A1A, A2C, A3E/F, A4G, A5N, A6A, A7I, A8J, A9K/G, A10N/L,
C3	A1B, A2D, A3F, A4H, A5M, A6A/B, A7I, A8J, A9K/G, A10N/L,

To indicate pattern entanglement and find out whether PDD is able to locate them in the synthetic data, the same sub-pattern is implanted in two different classes, indicating that the sub-pattern is common to and overlapping in both classes. Moreover, the sub-pattern also overlaps with other sub-patterns of the same or different classes. Hence, in a certain sense, patterns from different groups and even from the same entities are entangled. Though they are relating to different classes yet are difficult to separate, unless their associations with other AVs can be identified in AVA disentangled spaces. As our experimental results show, without disentanglement, traditional methods, such as Apriori [9] and HOPD (High-Order Pattern Discovery) [10] [11], produces far too many redundant/overlapping patterns (Figure 5) but fails to render succinct pattern clusters associating with the implanted classes. The objective of this experiment is to show how these embedded patterns associating with different classes are entangled yet can be identified and interpreted succinctly as disentangled patterns from a small set of AVA clusters in the disentangled space with no/least entanglement without relying on the input knowledge.

The results of Apriori [9] vary depending on the set value of support or confidence. For the threshold supp=20% and con=80%, it discovered 254 patterns associating with C2 and C3 but none with C1, since many of the C1 patterns overlap with C2 and C3 as partially shown in Figure 5(A). For support=10%, 4041 patterns associating with various classes were discovered. Many were associating with AV's containing in the attributes with random values (A11 to A15). When using Apriori, the threshold for support and confidence need to be set with class labels given. When HOPD [10], with significant level = 1.96 and the maximum order = 10, was applied, a large number (12,312) of patterns (up to order 10) were discovered. This number could be reduced by lowering the order of the patterns. Figure 5(B) shows the combinatorial nature of the patterns discovered. Though patterns are ranked, it is difficult to derive succinct interpretation. When pattern clustering [11] was applied, three pattern clusters were obtained (Figure 6). Three pattern clusters with 1084, 7758 and 3470 patterns were discovered. Each cluster contains patterns from associating with different or no class label. Each table shows the AV distributions of the attributes characterizing that cluster. The patterns are subsets of AV combinations in the cluster. The patterns were entangled in each cluster. Furthermore, both space and time complexity are high.

When PDD was applied to **R,** a much small set of patterns were obtained from the PG/SubPG in the disentangled spaces without relying on class information. PDD discovered only 43 statistically significant patterns, and 22 of them pertaining to distinct classes (i.e. C1, C2, C3) as shown in the

comprehensive PDDKB (Figure 7). The implanted patterns are discovered, and the original entangled patterns are disentangled in the result of PDD in different DSs which can be explicitly interpreted. In order to show patterns discovered from the AVA disentanglement spaces and subgroups in the space, the Disentangle Space Unit (DSU) is introduced. As patterns are discovered from AV-clusters from different DS*, the SubPG1 is referred as the Pattern sub-cluster 1, PG2 as the Pattern cluster 2. Hence, a pattern with a triple code DSU[3 2 1] represents the pattern is in the DS3, Pattern cluster2 and Pattern Sub cluster1.

In the Comprehensive PDDKB (Figure 7), the implanted patterns pertaining to C1 and C2 were found in PSG1 of DSU[1 1 1] in DS1 and PSG1 of DSU[2 1 1] in DS2 respectively in the AVA disentangled spaces. Patterns pertaining to C3 were found in both PG2 of DSU[1 2 1] and PG2 of DSU[2 2 1] respectively. Since PDD discovers patterns not based on classed labels, patterns pertaining to different classes can be found in the same DS if they share strong sub-pattern(s). For example, a pattern in DSU[2 2 1] consisting of AVAs [A4H, A5M] pertains to C1 was discovered in the same DS* contains other patterns pertaining to C3. This can be interpreted that the sub-pattern A5M and A7D is actually the entangled sub-pattern of C1 and C3 as shown in Table 2. However, the other sub-pattern of C1 consisting of [A3F, A4H, A5M] sharing A5M is common to a sub-pattern of C1 in DSU[1 1 1]. Nevertheless, each pattern, which may consist of sub-patterns found in other classes, is still associating with one distinct class. The pattern discovery results fully demonstrate the precise and explicit interpretability of PDD. Hence, in clinical application, PDD is able to discover a subset of signs-and-symptoms shared by different disease complexes explicitly and succinctly. Furthermore, unlike cases in Apriori (Figure 5(A)), no pattern discovered by PDD was associating with A10-A15 since they were random attributes containing only noise. This also demonstrate the superior noise avoidance capability of PDD over other ML methods. When checking the entities clustering result, all those entities labeled as C1 were clustered into EG within DSU[1 1 1] and all those labeled as C2 and C3 were clustered within DSU[2 1 1] and DSU[1 2 1] respectively.

Hence, this synthetic experimental result demonstrates the necessity and efficacy of the pattern disentanglement of PDD to relate concise/precise patterns to classes and the sources/causes as implanted through the experiment design. As the result shows, PDD renders a much small set of patterns (only 22) from the DS*s. They are succinctly and accurately associating with the implanted classes, while traditional methods, like Apriori [9] and HOPD [10] [11], usually produce an overwhelming

number of overlapping patterns with confused association due to their entanglement in the source environment particularly if the order of patterns specified is high.

Figure 5

(A) Pattern samples obtained by Apriori

Consequent	Antecedent	Support %	Confidence %
Class = C2	A5 = N	33.333	100
Class = C2	A4 = G	33.333	100
Class = C3	A1 = B	33.333	100
Class = C1	A7 = D	33.333	100
Class = C1	A8 = F	33.333	100
Class = C3	A2 = D	33.333	100
Class = C2	A5 = N and A4 = G	33.333	100
Class = C2	A5 = N and A2 = C	33.333	100
Class = C2	A5 = N and A1 = A	33.333	100
Class = C2	A5 = N and A7 = I	33.333	100
Class = C2	A5 = N and A8 = J	33.333	100
…	…	…	…
Class = C2	A14 = P and A5 = N	10.333	100
Class = C2	A14 = P and A4 = G	10.333	100
Class = C1	A13 = Q and A7 = D	11.333	100
…	…	…	…

(B) Pattern samples discovered by traditional Pattern Discovery Approach

Index	Residual	Prob.	Occurence	Order	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	Class
0	32.35	0.33	1000	5		C		H	M		D	F								
1	32.35	0.33	1000	5	B		F	H	M											C3
2	32.35	0.33	1000	5	B			H			I	J								C3
3	32.35	0.33	1000	5	B	D			M		I	J								
4	32.35	0.33	1000	5	A	C		H			D	F								
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
12310	1.96	0.05	160	6		C		H	M				G	N		P				
12311	1.96	0.05	160	6	A	C			M				G	N		P				

Figure 6 Pattern Clusters Obtained by Traditional Pattern Clustering Method

Pattern Cluster 1

A1		A2		A3		A4		A5		A6		A7		A8		A9		A10		A11		A12		A13		A14		A15		Class	
A	0.71	C	0.71	E	0.36	H	0.64	M	0.64	B	0.27	D	0.36	F	0.36	G	0.72	N	0.50	P	0.27	Q	0.35	P	0.37	P	0.29	Q	0.41	C1	0.36
B	0.29	D	0.29	F	0.64	G	0.36	N	0.36	A	0.73	I	0.64	J	0.64	K	0.28	L	0.50	Q	0.44	O	0.34	Q	0.30	Q	0.37	O	0.30	C2	0.36
																				O	0.28	P	0.31	O	0.32	O	0.34	P	0.30	C3	0.29

Pattern Cluster 2

A1		A2		A3		A4		A5		A6		A7		A8		A9		A10		A11		A12		A13		A14		A15		Class	
A	0.64	C	0.64	E	0.30	H	0.64	M	0.64	B	0.25	D	0.29	F	0.29	G	0.68	N	0.47	P	0.26	Q	0.31	P	0.36	P	0.31	Q	0.40	C1	0.29
B	0.36	D	0.36	F	0.70	G	0.36	N	0.36	A	0.75	I	0.71	J	0.71	K	0.32	L	0.53	Q	0.42	O	0.33	Q	0.30	Q	0.38	O	0.31	C2	0.36

Pattern Cluster 3

A1		A2		A3		A4		A5		A6		A7		A8		A9		A10		A11		A12		A13		A14		A15		Class	
A	0.64	C	0.64	E	0.33	H	0.71	M	0.71	B	0.30	D	0.36	F	0.36	G	0.72	N	0.50	P	0.27	Q	0.32	P	0.40	P	0.29	Q	0.36	C1	0.36
B	0.36	D	0.36	F	0.67	G	0.29	N	0.29	A	0.70	I	0.64	J	0.64	K	0.28	L	0.50	Q	0.42	O	0.36	Q	0.32	Q	0.38	O	0.31	C2	0.29

Figure 7 PDDKB obtained by PDD, both Summary PDDKB and Comprehensive PDDKB

Summary PDD Knowledge Base

DS	PG	SubPG				A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	Class
1	1	1				A	C	E	H	M	B	D	F	G							C1
1	2	1				B	D	F			A	I	J	K							C3/C2
1	2	2							G	N		I	J								
2	1	1				A	C	E	G	N	A	I	J								C2
2	2	1				B	D	F	H	M	B	D	F								C3/C1

Comprehensive PDD Knowledge Base

DS	PG	SubPG	Residual	Order	Occr.	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	Class
1	1	1	47.16	9	220	A	C	E		M	B	D	F	G							C1
1	1	1	69.78	9	460	A	C		H	M	B	D	F	G							C1
1	1	1	67.64	9	490	A	C	E	H	M		D	F	G							C1
1	1	1	48.34	9	220	A	C	E	H	M	B	D	F								C1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...							...
1	2	1	53.31	6	1000	B	D	F				I	J								C3
1	2	1	5.38	6	200			F			A	I	J	K							C2
1	2	1	44.24	7	450	B	D	F				I	J	K							C3
1	2	1	39.23	7	630	B	D	F			A	I	J								C3
1	2	1	35.67	8	310	B	D	F			A	I	J	K							C3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...							...
2	1	1	78.11	8	1000	A	C		G	N	A	I	J								C2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...							...
2	2	1	30.39	6	370	B	D	F		M	B										C3
2	2	1	20.68	6	370	B		F	H	M	B										C3
2	2	1	26.36	6	510			F	H	M		D	F								C1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...							...

4.2 Pattern Discovery Bias Avoidance Result on Datasets with Imbalanced Classes

4.2.1 Materials

Similarly, to validate the pattern discovery performance of PDD on imbalanced datasets, two imbalanced datasets are used.

Dataset 1: Imbalanced Synthetic Dataset. The synthetic dataset 1 is stochastically generated as a 2100 x 10 matrix with the first column as the class label and others as attributes with character values

from a uniform distribution. Then patterns of three different classes C_1, C_2, and C_3 are embedded for the first 6 attributes. For example, A1A and A2C represent character values A and C for Attribute A1 and A2. The patterns implanted in the data are summarized in Table 3. 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, A7, A8 and A9, the random characters from {"O", "P", "Q"} are inserted to make them as noise attributes. Moreover, this synthetic Dataset is one with imbalanced class distribution with 1000 entities pertaining to C2 and C3 each, and 100 entities pertaining to C1.

Table 3 Embedded Entangled Patterns for Dataset 4: Imbalanced Synthetic Dataset

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 - Imbalanced Clinical Dataset: 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 [37]. The attributes included are given in Figure 8. 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.

4.2.2 Experimental Result on Imbalanced Datasets

In this analysis, PDD is applied on both imbalanced synthetic and clinical datasets (Thoracic dataset). First, discovered patterns obtained through PDD, Apriori [38] (a typical frequent pattern mining method) and a HOPD [10] (our early work closely resembling the pattern discovery reported in [32] [16]) are compared. Figure 9 and Figure 10 show the pattern discovery result of PDD on the Synthetic and Thoracic data respectively. Figure 11 presents the comparison results of all these three methods.

From the pattern discovery result on the synthetic data, it is observed that:

Figure 8 Attribute Description of Thoracic Dataset

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)**

i. A small set of four AV-Clusters (Figure 9(A)) was discovered by PDD. From these clusters a comprehensive set of patterns (Figure 9(B)) were discovered. Each of these patterns associates with a distinct class or subgroup in a DS*.

ii. PDD discovered both the summarized patterns (Figure 9(A)) and the detailed patterns (Figure 9(B)). The summarized patterns are the unions of attribute values used for growing all comprehensive patterns (high-order patterns.)

iii. Figure 11 (A) shows the comparison results of the three methods in terms of their capability of discovering patterns from the rare classes as well as the number of patterns each method discovers. PDD is able to discover the pattern associating to the implanted patterns of the rare

class (Table 3) from AV Cluster 1 of DS2 (Figure 9(A)) whereas HOPD failed to discover due to the overwhelming number (770) of overlapping and entangled patterns (Figure 11(A)) as disclosed. Apriori, after fine-tuning the support and confidence, discovered only a few rare cases (Figure 11(A)) but also a large number of patterns (946 and 962). Note that PDD discovered only 9 patterns which correspond with the implanted patterns/sub-patterns (Table 3).

Furthermore, when comparing the implanted patterns with the pattern discovery result of PDD (Figure 9(C)), it is observed that all patterns (in colored bold fonts) discovered by PDD associates with correct classes in disentangled spaces except one in the last row of P2 as it has a sub-pattern (A2D, A5B, A6J) overlapping with a sub-patterns in C3 (Figure 9(C)). However, the values of the statistical residual of the overlapped patterns are obviously lower than the others from both C2 and C3. These results could be influenced by the implanted entangled sub-patterns [A5B, A6J], which is exactly the sub-patterns in C2 entangled with that in C3. This explains why these patterns are all found in AV Cluster 2 of DS2.

Figure 9(C) also shows high SR for the implanted patterns assigned with the correct classes and low SR for the entangled cases such as those shown in the last row of P2 and P3. In Figure 11(A), we observed that for both Apriori or HOPD, they discovered a large number of patterns which are most likely redundant and/or overlapping with one another. It is also noted that some of the discovered patterns are associating with class labels while some others are associating with AVs in the noise columns A7, A8 and A9.

From the pattern discovery result on the Thoracic dataset, similar phenomena as described in item (i) to (iii) as in the case on the synthetic data if replacing Figure 9 and Figure 11(A) with Figure 10 and Figure 11(B) respectively. Figure 10 (A) shows 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 (Figure 10(B)).

Figure 11(B) shows the comparison results of PDD and other two methods when applying to the Thoracic data. First, 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 huge. Apriori only outputs 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. Hence, the number of high order patterns 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, it should be examined 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 of *support* 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 PDD 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 and statistical/functional characteristics of both classes even when the class distribution is imbalanced.

With the capability to render a small, succinct and reliable set of patterns discovered from distinct sources, PDD is different from the existing model-based approach to rely on a priori knowledge and post processing.

In the clinical data analysis, in order not to spend significant involvement and efforts in precise planning before the analysis, domain users would like to perform prediction without requiring sophisticated handling [3]. PDD performs good results with good interpretability without knowing priori knowledge.

In reference [39], the authors presented the rules they discovered from the Thoracic dataset. For example, the rule "(PRE14 = OC14) => Risk1Yr = T" as presented in the reference [39] with the highest accuracy was also discovered in the result of PDD in PG1 of DS1. The experimental results show that PDD is possible to identify cases of higher risk of patient's death after surgery. In the result of PDD, when the value of attribute of PRE8 to PRE11 (Figure 8) are recognized as T, patients may have the higher risk of death after surgery, otherwise, patients have lower risk.

Figure 9 PDD Pattern Discovery Result from Synthetic 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) In the First and the Second DS* two AV-clusters were discovered

		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) Detailed Patterns associating with different classes were discovered from the above two AV-Clusters.

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) Comparison between implanted patterns and PDD's output.

Figure 10 Pattern Discovery Result of Thoracic Dataset using PDD

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) In both First and the Second Disentangled Space, two AV Clusters corresponding to Risk1=T and RISK1=F were discovered

		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) Detailed Patterns discovered from the above tow AV Clusters.

Figure 11 Comparison Result between PDD and Traditional Pattern Discovery Approaches on Imbalanced Datasets

Synthetic Data	Apriori		HOPD	PDD	
	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) Comparison of Pattern Discovery Result on Synthetic Dataset

Throcic Data	Apriori		HOPD	PDD	
	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) Comparison of Pattern Discovery Result on Thoracic Dataset

Chapter 5

Unsupervised Learning Based on PDD

5.1 Clustering using PDD

As mentioned in Chapter 3, the first set of output of PDD consists of the discovered AVAs and AV Clusters obtained from a small statistically significant set of DSs with maximum SR in their RSRV exceeding the threshold of statistical significance. Based on the AVs pertaining to the AV clusters, other outputs such as entity clusters and anomaly cases can be detected. The entity clusters can be obtained by maximizing the overlapping number of AVs shared by a group of entities and the AV clusters. The anomaly cases, which are defined as the entities not possessing AVAs pertaining to their labelled group (class) but to no class or other classes, can be obtained from their relation to AV Clusters found by PDD.

For relational datasets in a clinical setting, an “entity" in PDD represents a patient or a patient record. Thus, the Entity Clustering (EC) is the process to cluster the patients according to their characteristics discovered in the pattern discovery process without using class information. To state it formally, the Entity Clustering (EC) process in Disentangled Spaces is a process to assign each entity in **R** to an AV Cluster in certain DS* by maximizing the number of AVs that the entity share with the AV cluster, i.e. the number of AVs of that entity found in that AV cluster. Its pseudo-code is given in Algorithm 3.

Algorithm 3: Entity Clustering

Input:
$\boldsymbol{AVCluster} = \{AVCluster_1, \ldots AVCluster_{n+1}\}$
$\boldsymbol{R}\ (A^i = \{A_n^i | A_n^i \epsilon \{e_{n1}, e_{n2}, \ldots e_{nI_n}\}\}$
Output: $EC = \{EC_1, \ldots EC_{n+1}\}$
Procedure ***EntityClustering***

Begin
  For each entity, A^i, in R
    For each $AVCluster_k$
      $Sharing = A^i \cap AVCluster_k$
    End
  Assign A^i into EC_k by maximizing *sharing*
  End
  Return $EC = \{EC_1, \ldots EC_{n+1}\},\ EC_k = \{A^i | A^i = [A_1^i, \ldots A_N^i],\ A_n^i \epsilon \{e_{n1}, e_{n2}, \ldots e_{nI_n}\}\}$
End

5.2 Materials

To show that PDD can relate pattern/entity clusters with class/functionality, two clinical and one pathological datasets are used.

Dataset 1: Heart Disease Data Set (Heart): Heart Disease [40] dataset is a health care benchmark dataset from UCI repository [41], which contains 270 clinical records with 13 mixed-mode attributes in two possible classes: Absence or Presence (of heart disease).

Dataset 2: Breast Cancer Wisconsin Data Set (Cancer): The Breast Cancer Wisconsin dataset [42] is a health care benchmark dataset taken from UCI repository [41], which is a classical dataset with 682 cases for discriminating the instances of two possible classes: Benign (distribution=65.5%) and Malignant (distribution=34.5%). Figure 12 shows the descriptions of the attributes in Heart Disease dataset and Breast Cancer dataset.

Figure 12 Attributes in Heart and Cancer Datasets

Attributes in Heart Data Set
1) **age**
2) **Sex**
3) **cpt**: chest pain type (4 values)
4) **rbp**: resting blood pressure
5) **sc**: serum cholestoral in mg/dl
6) **fbs**: fasting blood sugar > 120 mg/dl
7) **rer**: resting ECG results (0,1,2)
8) **mhra**: maximum heart rate achieved
9) **eia**: exercise induced angi
10) **oldpeak**: ST depression (exercise/rest)
11) **spess**: slope of peak exercise ST segment
12) **nmvc**: number of major vessels (0-3)
13) **thal**: 3=normal; 6=fixed defect

Class: Absence/Presence of Heart Disease

Attributes in Breast Cancer Data Set:
1. **Clump Thickness**: 1 - 10
2. **Uniformity of Cell Size**: 1 - 10
3. **Uniformity of Cell Shape**: 1 - 10
4. **Marginal Adhesion**: 1 - 10
5. **Single Epithelial Cell Size**: 1 - 10
6. **Bare Nuclei**: 1 - 10
7. **Bland Chromatin**: 1 - 10
8. **Normal Nucleoli**: 1 - 10
9. **Mitoses**: 1 - 10

Class: (2 for benign, 4 for malignant)

5.3 Entity Clustering

Although the process of clustering individuals does not require class label information, the performance of the clustering results can be evaluated by two statistical measures using the presumed class labels as ground truth. One was "accuracy" which was defined as

$$\frac{\textbf{\textit{The number of patients labeled correctly}}}{\textbf{\textit{The total number of patients}}}$$

The accuracy reflected the quality of the discovered clusters [46]. The more commonly used measure was F-measure [47], which was calculated using True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) by the equation:

$$\boldsymbol{F-measure} = \frac{\boldsymbol{2TP}}{\boldsymbol{2TP + FN + FP}}$$

The result of PDD is compared with K-means. SPSS Modeler 14.1 was used for K-means clustering which accepted a mix-modal data types, including both categorical and numerical values in a dataset.

5.3.1 **Experimental Result on Clinical Datasets**

In this section, two benchmark datasets, Heart Disease data set (Heart) and Breast Cancer data set (Cancer) were used for clustering. Figure 13 and Figure 14 show the comparison results of clustering for the Heart and the Cancer datasets respectively.

For Heart Disease, Figure 13 shows that PDD (F-measure=0.83, Accuracy=82.87%) outperforms K-Means on both original numerical (F-measure=0.59, Accuracy=59.26%) and discretized datasets (F-measure=0.81, Accuracy=81.48%) in F-measure and Accuracy. For Cancer, Figure 14 shows the results of Accuracy and F-measure of PDD vs K-Means on the discretized datasets are closer since this dataset contains less noise. But the leverage is that PDD could reveal all the patterns in the Entity Clusters while K-Means could not. It opens the door to visualize patterns in clusters formed.

Figure 13 The comparison of entity clustering result of K-means (on numerical data (N) and discretized data (D)) and PDD on Heart Disease Dataset

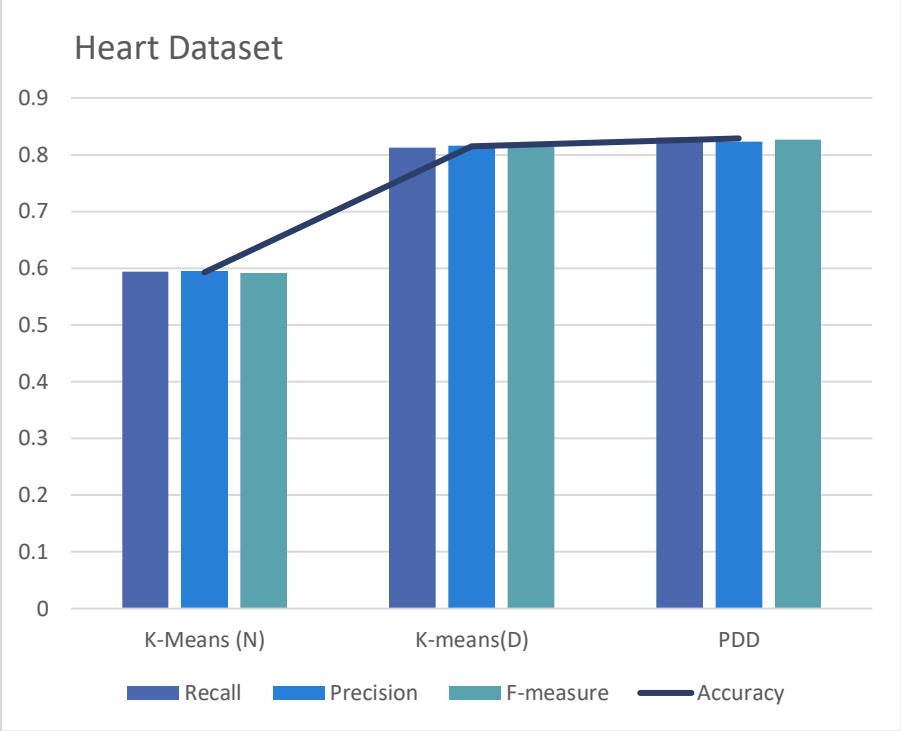
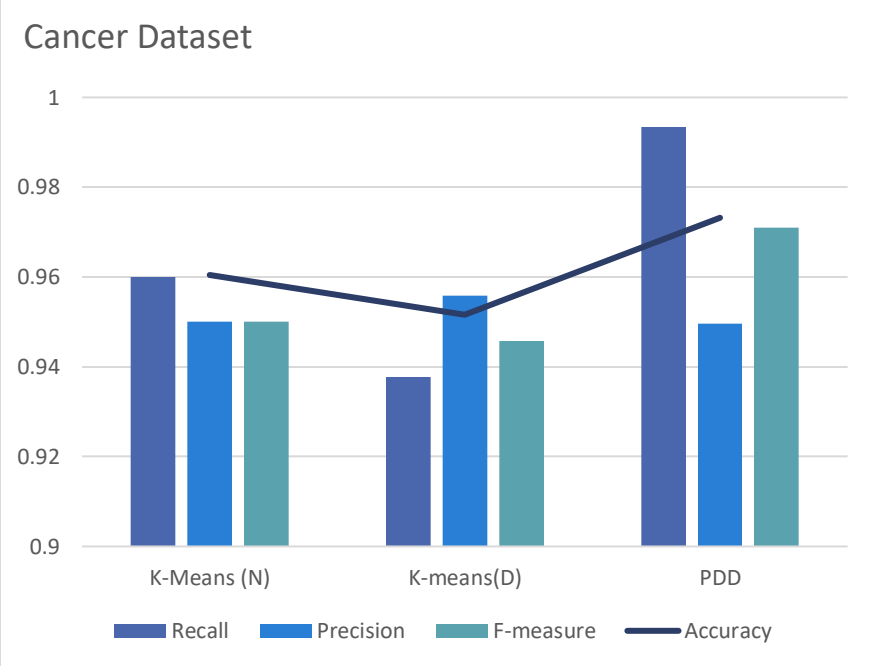


Figure 14 The comparison of entity clustering result of K-means (on numerical data (N) and discretized data (D)) and PDD on Breast Cancer Dataset.



5.4 Anomaly Detection

5.4.1 **Experimental Result on Clinical Datasets**

To demonstrate PDD's ability to identify anomalies and improve the classification accuracy if they are identified and removed from **R** before training and classification, the Heart Disease dataset and Cancer dataset were used first.

After entity clustering, the entities (patients) are clustered into different groups. The anomaly check can be conducted. The *anomalies*, which are defined as the entities not possessing AVAs pertaining to their labelled group (class) but to non-class or other classes, can be obtained from their relation to the AV Clusters found by PDD. The common patterns for PD eligibility and anomaly cases were reported in the following section. These anomaly entities may arise from: 1) outliers in the given dataset; or 2) some entities may correspond to an anomaly case or an early stage of disease although being labeled as "healthy".

As mentioned in chapter 3, the PDDKB contains summarized PDDKB and comprehensive PDDKB. The summarized PDDKB contains three sections.

1) In the DS section, each DS Unit (DSU), identified by a triple code made up of the index of the DS*, pattern cluster and sub-cluster, represents the disentangled AVA source from which the pattern is discovered. For example, a pattern would associate with a DSU [3 2 1] if it is discovered in the sub-cluster 1 of the pattern cluster 2 of DS3.

2) The summary PDDKB summarizes all the patterns as a super-pattern (the union of the patterns) in each DSU denoted by their triple code.

3) In the entity section, each column represents an individual with a distinct EID and its class label (if given). The numeral on a column and a row represents the number of patterns which the entity on that column possesses in the specific DSU on that row. For example, in Figure 17, the numeral 8 for entity E37 denotes that it contains 8 patterns in the DSU [1 1 2] as shown in the comprehensive PDDKB.

In the Comprehensive PDDKB, all the patterns/AV-clusters possessed by an entity are listed on that column. Hence, PDDKB encompasses all the integrated deep knowledge discovered from **R.** Comprehensive PDDKB also contains three sections same as those described in the summarized

PDDKB except that each row contains just a one discovered pattern from the DSU.

In the summary section, the summary patterns summarize the AV clusters (or pattern clusters) listed in the DSU in the Comprehensive PDDKB. For instance, the AVs in the first row represents the union of all AV clusters found in the DSU in the Comprehensive PDDKB. Each of them links to a list of individual entities (denoted by '1') in the Entity Section (Entities) where each column represents an entity with EID and class label (if given). In the Summary PDDKB, the numeral on each column (like 8 associating with E37 in Figure 17) denotes the number of patterns/AV-clusters discovered from the DSU [1 1 2]. In the Comprehensive PDDKB, on the same column, a numeral of "1" is displayed on the row containing a special AV cluster (or pattern) that the entity possesses.

The result of anomaly detection of Heart dataset and Cancer dataset are presented in Figure 15 and Figure 17 respectively. As Figure 15 shows, entities E122 and E131 are rare cases since they are labelled as "Absence" but possess patterns pertaining to the "Presence" group.

Figure 15 Summary PDDKB and Comprehensive PDDKB Obtained from Heart Dataset.

Summary PDD Knowledge Base

DS				Summary Patterns													Entities (E)									
Disease Complex/Class				sign/symptoms/lab tests													Absence					Presence				
DS	PG	SubPG	class	age	sex	cpt	rbp	sc	fbs	rer	mhra	eia	oldpeak	spess	nmvc	thal	1	2	...	122	131	151	152	...	269	270
1	1	1	Absence	[29 51]	F	2;3					[162 202]	0	[0 0.1]	1	[0 1]	3	33	6	...					...		
1	2	1	Presence	[59 77]	M	4					[71 143]	1	[1.4 6.2]	2	[1 3]	7			...	1	4	45	3	...	5	66
2	1	1	Presence			4								2	[1 3]	7			...			7	1	...	1	7
2	2	1	Absence											1	[0 1]	3	5	2	...					...		

PDD disentangles the dataset into two DSs. In DS1, two Pattern Groups are discovered for Absence and Presence. In DS2, two pattern groups are discovered with low-order patterns.

Dataset contains 13 mixed-mode attributes (i.e. Real, Ordered, Binary, Nominal). Each row represent a summary patterns (Attribute Cluster). e.g. The first row represents in DS1 and PG1, the attribute cluster contains 11 attribute values (class=Absence; age=[29 51]; sex=F; cpt=2/3; mhra=[162 202]; ... ... thal=3.)

270 entities in Heart Disease data set. Entities 1-150: Absence; 151-270: Presence. The value in the block, such as 33, means there are 33 high-order patterns are grown from the first AV Cluster covered by the first entity. Each of these patterns will be displayed in the comprehensive PDDKB

Comprehensive PDD Knowledge Base

DS	PG	SubPG	Residual	Order	Occr.	class	age	sex	cpt	rbp	sc	fbs	rer	mhra	eia	oldpeak	spess	nmvc	thal	1	2	...	150	151	...	269	270
1	1	1	6.87	3	103	Absence									0				3	1	1	...			...		
1	1	1	9.18	4	73	Absence									0		1		3	1	1	...			...		
1	1	1	9.3	4	57	Absence		F							0				3		1	...			...		
1	1	1	7.37	4	46	Absence			3						0				3	1		...			...		
1	1	1	9.43	5	27	Absence		F	3						0				3			...			...		
1	1	1	9.94	5	43	Absence		F							0			[0 1]	3			...			...		
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1	2	1	6.88	3	68	Presence		M											7			...			...	1	
1	2	1	4.77	3	64	Presence		M									2					...		1	...		1
1	2	1	4.36	3	38	Presence	[59 77]										2					...		1	...		1
1	2	1	6.57	3	58	Presence			4								2					...		1	...		1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

For the Cancer dataset, to exemplify PDD's capability to discover patterns for small/rare classes and discriminate biases/anomalies [49], two small transition groups are inserted into the dataset -- Transition1 and Transition2 (with 30 samples each, 4% of the whole data). They were stochastically generated with transitional AVs from Benign to Malignant to mimic the early stage of cancer. Figure

16 gives the quantized AVs of the transition groups. The yellow and green blocks are the majority patterns from Benign and Malignant classes respectively. The first 682 samples were taken from the original data and those from 683-712 and 713-742 were taken from Transition1 and Transition2 respectively. These small transition groups, if spotted, may help to detect the progression of cancer from early to late stage [50].

Figure 16 The inserted patterns for the rare case groups of Cancer Dataset. (Data quantization put each AV for each group in the same intervals.)

Class	Clump thickness	Cell Size	Cell Shape	Marginal Adhesion	Signle Cell size	Bare Nuclei	Bland	Nucleoli	Mitoses
Transition 1	[1 3]	[3 10]	[3 10]	[3 10]	[1 2]	[1 3]	[1 3]	[1 2]	Random
Transition 2	[5 10]	[1 3]	[1 3]	[1 3]	[3 10]	[3 10]	[3 10]	[2 10]	Random

In the PDDDB given in Figure 17, all 743 entities were included though not totally shown in the list. Most of them associate with correct DSU/class labels. However, PDD unveils some outliers such as E36 shaded in yellow. An outlier is an entity that does not possess a pattern according to a prescribed statistic threshold, whereas a rare case is one which possesses patterns of classes not as labeled in **R** (green shaded boxes) (Figure 17). For example, E407 and E422 were labeled as Benign but both possess patterns discovered as associating with Malignant with none in the Benign. Similarly, E462 was labeled Malignant, but possesses only patterns in the Benign. In healthcare, it is crucial if rare patients could be spotted earlier before therapy and treatment because they may be misdiagnosed patients or the early cases of the disease.

Once the PPDKB is completed, simple algorithms can be used to accomplish various ML tasks and naturally allows integrated analytics, interpretation, knowledge tracking and organization to fulfil the goals of precise data analytics.

Figure 17 Summary PDDKB and Comprehensive PDDKB Obtained from Cancer Dataset.

Summary PDD Knowledge Base

DS				Summary Patterns									Entities (E)														
													Benign						Malignant				Transition1			Transition2		
DS	PG	SubPG	class	Clump thickness	Cell Size	Cell Shape	Marginal Adhesion	Signle Cell size	Bare Nuclei	Bland	Nucleoli	Mitoses	1	...	36	37	407	422	444	...	462	463	683	...	712	713	...	743
1	1	1	Benign	[1 3]	[1 3]	[1 4]	[1 3]	[2 3]	[1 4]	[1 2]/[2 3]	[1 2]		18	...		1					1		1	...	1			
1	1	2	Benign	[3 5]	[1 3]	[1 4]	[1 3]	[2 3]	[1 4]	[2 3]	[1 2]					8													
1	2	1	Malignant	[5 10]	[3 10]	[4 10]	[3 10]	[3 10]	[4 10]	[3 10]	[2 10]						1	1	1			2				1	...	1
1	2	2	Transition2					[3 10]																		1	...	1
2	1	1	Transition1	[1 3]	[3 10]	[4 10]	[3 10]	[1 2]	[1 4]	[1 2]	[1 2]		1	...		1							4	...	4			
2	2	1	Transition2			[1 4]	[1 3]			[3 10]										...		1				1	...	1

PDD disentangle the data into three DSs. In DS1, Benign and Malignant are discovered in two opposite clusters; Transition2 is discovered in the same cluster with Malignant but in different Sub-cluster.

The data set contains 9 attributes transforming into discrete value in PDD result. Each row represents an attribute cluster or called summary pattern. e.g. The first row represents in the sub- pattern cluster 1 (SubPG=1) of pattern cluster 1 (PG=1) in disentangled space 1 (DS=1), the patterns are grows from the attribute cluster containing 10 attribute values (e.g. class=Benign; clump thickness = [1 3]; .....)

There are 742 entities in total, the first 682 entities are from original observations and the last 60 entities are generated rare cases. Benign: 1-443; Malignant: 444-682 Transition 1: 683-712; Transition 2: 713-743 The value in the block, such as 18 in the first block, means there are 18 high-order patterns are growth from the first AC that can be covered by the first entity.

Comprehensive PDD Knowledge Base

DS	PG	SubPG	Residual	Order	Occr.	class	Clump thickness	Cell Size	Cell Shape	Marginal Adhesion	Signle Cell size	Bare Nuclei	Bland	Nucleoli	Mitoses	1	2	3	4	5	6	7	8	9	...
1	1	1	14.74	4	163		[1 3]	[1 3]		[1 3]		[1 4]				1		1	1			1	1	1	...
1	1	1	22.75	4	358			[1 3]	[1 4]	[1 3]		[1 4]				1	1	1	1	1	1	1	1	1	...
1	1	1	25.89	4	368	Benign		[1 3]	[1 4]	[1 3]						1	1	1	1	1	1	1	1	1	...
1	1	1	21.93	4	314				[1 4]	[1 3]	[2 3]	[1 4]				1	1	1		1	1	1	1		...
1	1	1	22.66	4	305			[1 3]		[1 3]	[2 3]	[1 4]				1	1	1		1	1	1	1		...
1	1	1	24.85	4	377	Benign			[1 4]	[1 3]		[1 4]				1	1	1	1	1	1	1	1	1	...
1	1	1	15.45	4	93		[1 3]					[1 4]	[1 2]	[1 2]		1		1	1			1	1	1	...
1	1	1	22.55	5	166	Benign	[1 3]		[1 4]			[1 4]		[1 2]		1		1	1			1	1	1	...
1	1	1	23.03	5	167	Benign	[1 3]		[1 4]	[1 3]				[1 2]		1		1	1			1	1	1	...
1	1	1	25.99	6	130		[1 3]	[1 3]	[1 4]	[1 3]	[2 3]	[1 4]				1		1				1	1		...
1	1	1	27.45	6	132		[1 3]	[1 3]	[1 4]	[1 3]	[2 3]			[1 2]		1		1				1	1		...
1	1	1	19.29	6	55	Benign	[1 3]	[1 3]	[1 4]		[2 3]		[1 2]			1		1				1	1		...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

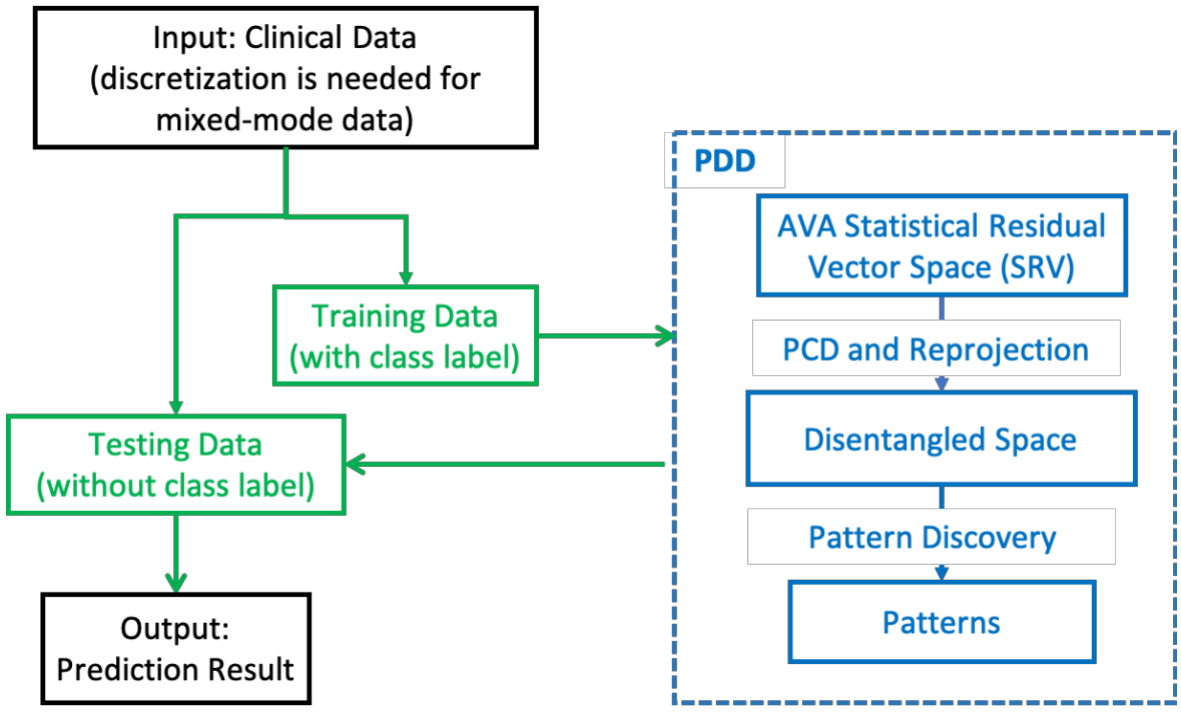
Chapter 6

Supervised Learning Based on PDD

6.1 Classification using PDD

As discussed in Chapter 3, the patterns discovered from the entire dataset can be used for experts' results interpretation. 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. For supervised learning, to evaluate the performance of the learner, patterns discovered from the training data with class labels given can be used for predicting the testing data. Then, the summarized patterns associated with a specific class discovered from the testing data are used to predict whether the entity belongs to that class or not. Figure 18 provides a schematic overview of the classification process using PDD.

Figure 18 Overview of Classification Process



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, the Weight

of Evidence [51] [31] of all the AVs in the summarized patterns are used to determine whether the class label for E_i, $C(E_i)$, will have more weight than that for predicting it as pertaining to other classes.

6.2 Materials

One of the novel components in the supervised learning process of PDD is the identification of outliers and rare cases. Due to the reasons for anomalies cannot be confirmed, those anomaly cases are removed before training process. Since if the anomalies are mislabeled cases or outliers, the result of prediction may be impacted.

Three clinical datasets were used to demonstrate PDD's ability to identify anomalies and improve the classification accuracy if they are identified and removed from **R** before training and classification. The first two datasets Heart Disease Dataset and Breast Cancer Dataset are described d in chapter 5. In order to demonstrate the efficacy of the performance of PDD for classifying imbalanced dataset, the thoracic dataset is used.

Imbalanced Clinical Dataset: thoracic dataset. This 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 [37]. The attributes included are given in Figure 19. 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.

6.3 Result

6.3.1 **Experimental Result on Clinical Datasets**

In this section, the comparison of classification results between PDD and other classification methods are provided for two clinical datasets, Heart Disease and Breast Cancer. The comparison results on the imbalanced thoracic datasets are reported in the next section. In all these experiments, 80% of the available data for each class was selected randomly as training data and the 20% remaining was retained as testing data. The classification accuracy results of PDD are compared with those from support vector machine (SVM) and artificial neural network (ANN) [52], using the original dataset and that after the

outliers and rare entities are removed. The experimental runs were iterated 10 folds to calculate the average classification accuracy for performance assessment and comparison.

To show that PDD can identify such distinct rare entities, for Heart Disease Dataset, all the abnormal entities and outliers are removed to produce a clean dataset which contains "Absence" entities, E1 to E130 and "Presence" entities, E131 to E237. Figure 20 shows the comparison results of classification accuracy between PDD and other algorithms. After the removal of anomalies, the classification results using different algorithms are improved approximately 10%.

Figure 19 Attribute Description of Thoracic Dataset

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)**

Similarly, Figure 21 shows the classification accuracy with variance by different algorithms for Breast Cancer dataset. There is 2-3% and 10% improvement on these datasets respectively after removing

outliers and rare entities identified by PDD. Actually, it is not meaningful if the outliers and rare entities are taken into the training, so the accuracy should be assessed using the ground truth when these entities are removed from the training set. The results show the removal is able to enhance the accuracy of all methods. As shown in the Breast Cancer Case (Figure 21), SVM, due to is strong discriminative ability, has superior performance than that of ANN and PDD though only about 1% higher. When the outliers and rare cases are removed, though its accuracy is improved 1% yet it is 1% lag behind that achieved by PDD. In this study, it is shown that PDD not only possesses the throughput/output transparency and interpretability, but also superior in classification accuracy especially when the outliers/rare entities are identified and removed while at the same time out-performs others much more in the case when the data size for different classes is imbalanced as shown in the next section. This is important for disease diagnosis since outliers not having significant disease associations and mislabeled in the training dataset can lower the diagnostic accuracy [2] [5].

Figure 20 Comparison of Classification Accuracy Result between Original Dataset and Dataset after Removing Anomalies on Heart Disease Dataset

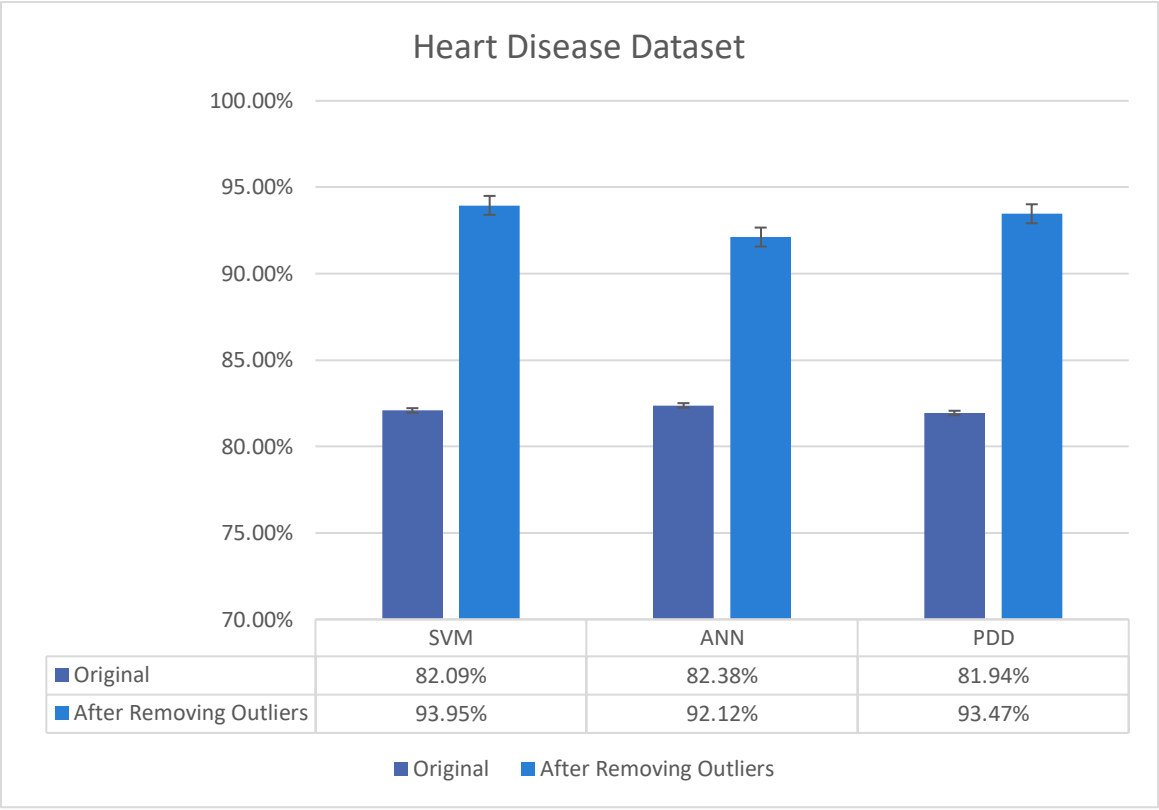
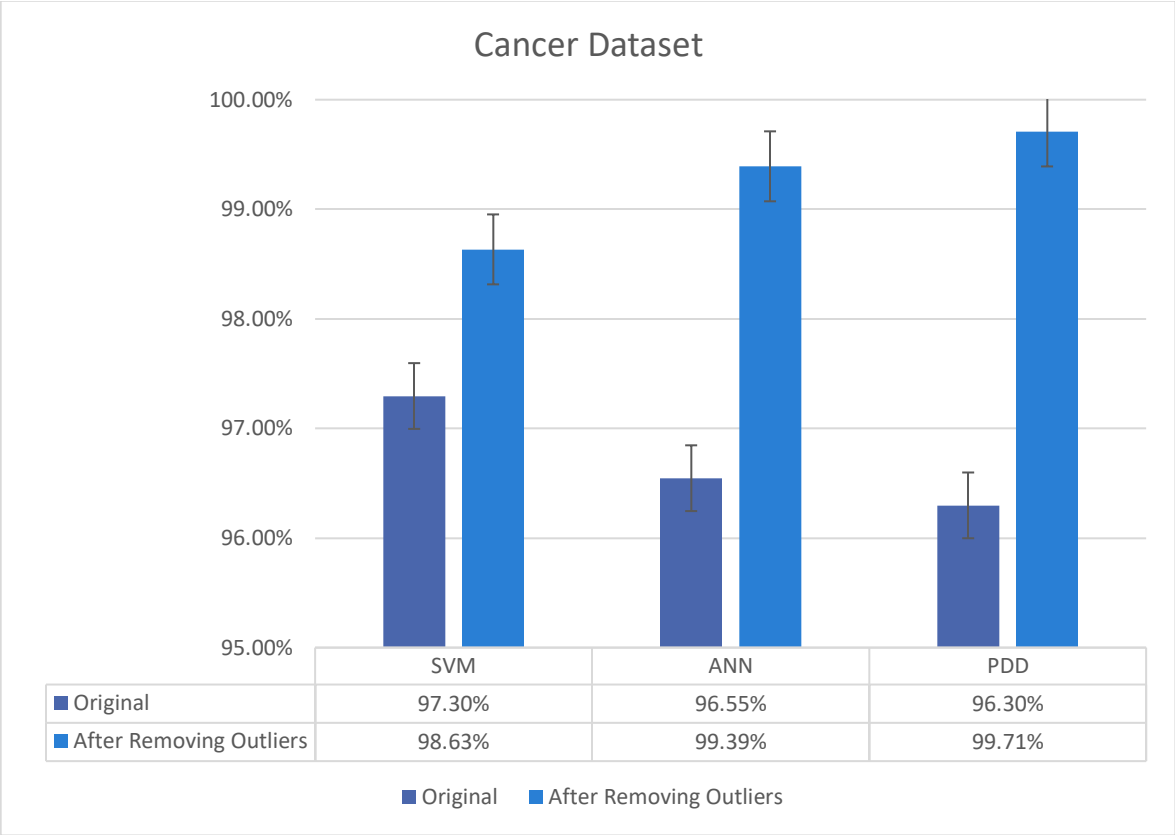


Figure 21 Comparison of Classification Accuracy Result between Original Dataset and Dataset after Removing Outliers on Breast Cancer Data Set



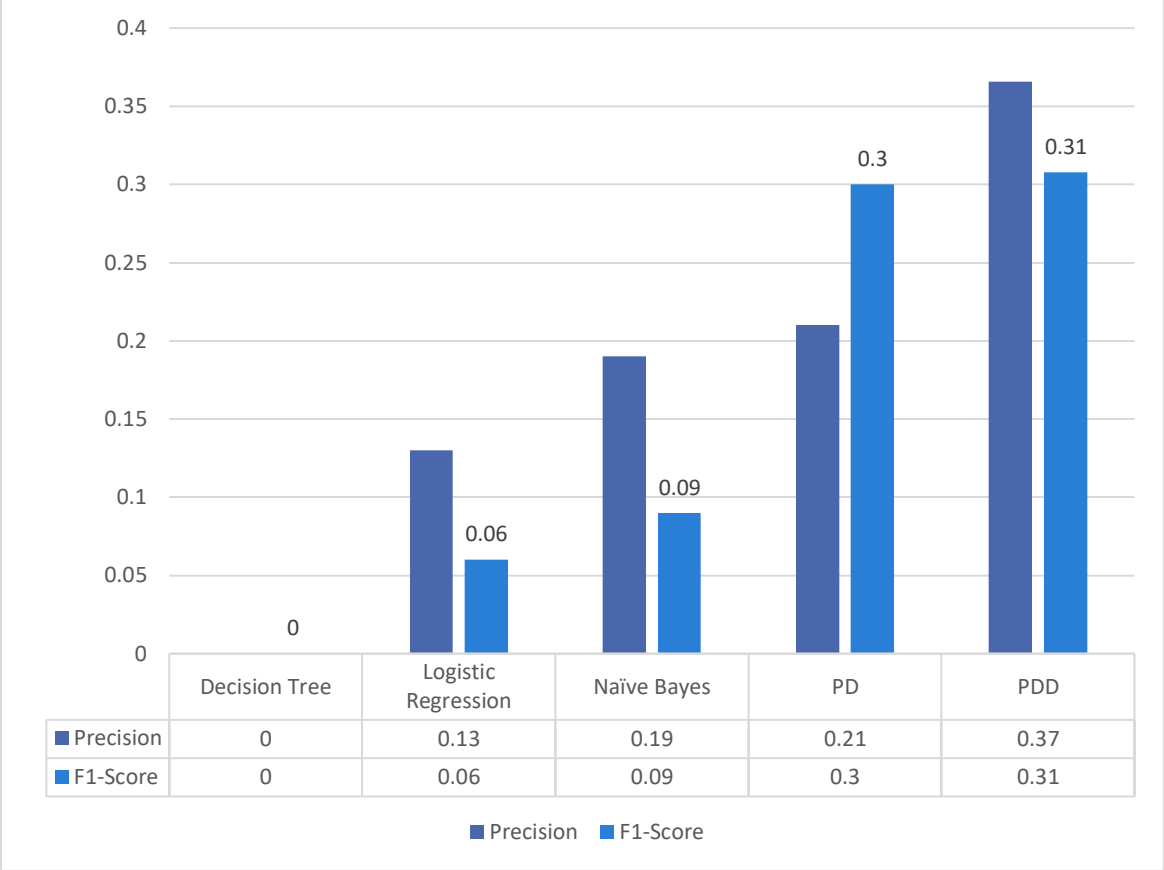
6.3.2 **Experimental Result on the Imbalanced Dataset**

In this section, to validate the performance of PDD for classification on imbalanced class, the prediction results of diagnostic outcomes of the Thoracic dataset with imbalance class distribution are provided. 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 were obtained and shown in Figure 22. The comparison results showed that PDD outperformed the other two classification methods. When comparing with a

recent pattern discovery method PD, PDD outperformed PD in both precision and F-measure. The results on the same dataset are taken from the work reported in [3]. The PD method [3] acquired lower precision rate than that of PDD, but a F1-Score of 0.3±0.01 which is close to that obtained by PDD (F1-Score=0.31±0.02). It is also noted that Decision Tree misclassified all the test cases since it did not discover any rule for the cases with Risk1Y=T. The experimental result on clinical data with high imbalanced class ratios shows that PDD does have a better interpretability and prediction performance for minority target. Due to the subtlety of the risk factors in the Thoracic data, though the precision and F1-score obtained by PDD are superior to those of the other methods, the ratios are still low. However, it shows that even for this type of fuzzy imbalanced data, PDD did pick up some subtle factors where other methods fail.

As the pattern discovery result on imbalanced data shown in chapter 4, PDD renders superior prediction performance and interpretability since it produces and uses much smaller set of succinct disentangled patterns. All the result 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 [4] as well as the problem of imbalanced class [2] [4] [5] [6]. As a clinical data analysis tool on relational data, it has a significant advantage over 'black box' ML algorithms as the outputs of cPDD is both interpretable and transparent, the two major challenges of interpretability and applicability [53] confronting ML today. Hence, PDD brings data analytics to clinical experts for direct interpretation of the discovered results to enhance their insight and understanding with statistical and rational accountability.

Figure 22 Average Classification Result from 20-times 10-fold Cross-Validation on Thoracic Dataset



Chapter 7

Conclusion

In this study, the proposed PDD, with a novel theoretic concept and effective and efficient algorithm design, plays a significant role in relational data analysis, especially the clinical data analysis. It is the first AI system to discover patterns from AVA disentangled spaces and attain deep knowledge using an all-in-one interpretable integrated knowledge base obtained by the system. It represents the results in a unified representation interlinking the sources, patterns and entities together to enhance accuracy, avoid biases and render interpretability for various ML tasks. Since the results of PDD obtained are robust, explicit and displayable for experts' interpretation, question-answering and knowledge base construction, it overcomes the limitations of current ML methods when confronting bias, rare groups, anomalies [2] [4] [5] [6] and lack of transparency [4]. PDD is an effective and credible method to render empirical evidence from relational data [3].

In data analytics, PDD discovers patterns while simultaneously assembling them into pattern/entity clusters to support decision-making and further knowledge exploration and organization, solving both supervised and unsupervised learning problems in ML. Furthermore, its use of Address Table and Entity ID to assist pattern discovery is a novel time-complexity reduction strategy versus exhaustive search in high dimensional feature spaces or manifolds [11] [10].

In addition, PDD has a significant advantage over 'blackbox' ML algorithms as it overcomes major hurdles --- interpretability, creditability and applicability --- in ML [53]. PDD outputs a statistical supported comprehensive and interpretable unified knowledge representation (PDDKB) containing a few smaller sets of distinct and explicit patterns/pattern-clusters related to different functional sources, interlinking patterns, source environments and individual entities for medical applications.

In Chapter 4, PDD shows the discovered patterns in disentangled spaces based on intrinsic statistically significant AVAs, even for imbalanced clinical data. It does not require explicit prior knowledge, which is often hard to get/justified. Then, In Chapter 5, it demonstrates that patterns discovered from the disentangled spaces are naturally better separated to produce more accurate results in pattern/entity clustering instead of relying on similarities. Since these clusters are more specific and less affected by the class/group size or biases, PDD is apt to solve the imbalanced class [6] [54] and anomalies problem [2], and can even go deeper to attain precise solutions for the rare/imbalanced cases. In Chapter 6, the

supervised learning result shows that PDD renders robust and credible solutions with high accuracy and interpretability, even for the clinical practice with imbalanced targeted groups.

PDD furnishes clinical/statistical support, linking diagnostic patterns to the etiological origins and individual patients with evidence explicitly displayable to medical professional, allowing the relevant experts or doctors to make further examination, testing, assessment and therapeutic decisions. Hence, it can contribute significantly to early disease prediction/diagnosis, therapeutic treatment, and prognosis evaluation of various conditions, particularly for depression [55], complex neuropsychiatric disorders such as Autism Spectrum Disorders [56] and stroke [57]. PDD can provide answers and interpretations to clinical questions/problems, which can be communicated to patients and/or healthcare helpers by physicians [2].

As future work, from the perspective of technology, the performance of PDD can be improved in the following two aspects. Firstly, different strategies and levels of discretization may impact the interpreting and prediction results of PDD. So, more experiments will be implemented to prove which discretization methods will be followed, machine learning methods applied automatically, or the clinical suggestions. Secondly, the traditional associative classification was applied for prediction. In order to improve the prediction performance using discovered patterns, more advanced machine learning strategies, such as ensemble classification methods and boosting associative classification method can be applied.

In addition, from the perspective of application, PDD can also be applied for proteomic and genetic medical studies. In proteomic, PDD can reveal imbalanced taxonomic classes (rare mutants) and subgroup characteristics of conserved functional domains, attaining accurate and explicit predictive analytic results without relying on prior knowledge. In addition, PDD will be extend to apply to unstructured data (e.g. text and sequences) [14] [58] by extracting AVAs directly from them as shown in our early work [59], which allow the patients' medical records are used as input data. Moreover, for performance improvement, parallel computing strategy will be introduced to handle bigger data and further speed up the computational time.

In conclusion, PDD can bridge the 'AI chasm'— the gap between creating a scientifically sound algorithm and its application to real-world problems [60]. It will play an important role in empirical and data sciences as it brings AI closer to experts with insight and accountability, meeting the scientific, economic, legal and social challenges for AI in healthcare and data analytics for the years to come.

Bibliography

[1] P. Voosen, "How AI detectives are cracking open the black box of deep learning," *Science,* 2017.

[2] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature medicine,* vol. 25, no. 1, pp. 44-56, 2019.

[3] T. 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.

[4] W. Samek, T. Wiegand and K. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296,* 2017.

[5] C. Aggarwal and S. Sathe, "Bias Reduction in Outlier Ensembles: The Guessing Game," in *Outlier Ensembles*, Springer, 2017.

[6] 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.

[7] Y. Sun, A. K. Wong and M. Kamel, "Classification of Imbalanced Data: A Review," *International Journal of Pattern Recognition and Artificial Intelligence,* vol. 23, no. 04, pp. 687-719, 2009.

[8] S. 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.

[9] C. C. Aggarwal and J. Han, Frequent pattern mining, Springer, 2014.

[10] A. 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.

[11] A. 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.

[12] 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.

[13] J. Cheng, Y. Ke and W. Ng, "\delta-Tolerance Closed Frequent Itemsets," in *Data Mining, 2006. ICDM'06. Sixth International Conference on. IEEE*, 2006.

[14] 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.

[15] A. 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.

[16] 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.

[17] P. Zhou, A. K. Wong and Z. A. Butt, "Pattern to Knowledge: Discovering Deep Knowledge from Relational Data by Pattern Discovery and Disentanglement," *Scientific Report,* Submitted.

[18] P. Shah , F. Kendall, S. Khozin, R. Goosen, J. Hu, J. Laramie, M. Ringel and N. Schork, "Artificial intelligence and machine learning in clinical development: a translational perspective.," *NPJ digital medicine,* vol. 2, no. 1, pp. 1-5, 2019.

[19] E. Gawehn, J. Hiss and G. Schneider, "Deep learning in drug discovery," *Molecular informatics,* vol. 35, no. 1, pp. 3-14, 2016.

[20] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer and S. Zhao, "Applications of machine learning in drug discovery and development," *Nature Reviews Drug Discovery,* vol. 18, no. 6, pp. 463-477, 2019.

[21] A. Esteva, B. Kuprel, R. Novoa, J. Ko, S. Swetter, H. Blau and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature,* vol. 542, no. 7639, pp. 115-118, 2017.

[22] V. Gulshan, L. Peng, M. Coram, M. Stumpe, D. Wu, A. Narayanaswamy and et.al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA,* vol. 316, no. 22, pp. 2402-2410., 2016.

[23] B. Alipanahi, A. Delong, M. Weirauch and B. Frey, "Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning," *Nature biotechnology,* vol. 33, no. 8, pp. 831-838, 2015.

[24] A. Rajkomar, E. Oren, K. Chen, A. Dai, N. Hajaj, M. Hardt and et.al., "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine,* vol. 1, no. 1, p. 18, 2018.

[25] A. Holzinger, C. Biemann, C. Pattichis and D. Kell, "What do we need to build explainable AI systems for the medical domain?," *arXiv preprint arXiv:1712.09923.,* 2017.

[26] S. Thonekaboni, S. Joshi, M. McCradden and A. Goldenberg, "What clinicians want: contextualizing explainable machine learning for clinical end use," *arXiv preprint arXiv:1905.05134,* 2019.

[27] D. Chen, S. Liu, P. Kingsbury, S. Sohn, C. B. Storlie, E. B. Habermann, J. M. Naessens, D. W. Larson and H. Liu, "Deep learning and alternative learning strategies for retrospective real-world clinical data," *npj Digital Medicine,* vol. 2, no. 1, p. 43, 2019.

[28] D. Castelvecchi, "Can we open the black box of AI?," *Nature News,* vol. 538, no. 7623, p. 20, 2016.

[29] 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.

[30] 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.

[31] N. Abdelhamid and F. Thabtah, "Associative classification approaches: review and comparison," *Journal of Information & Knowledge Management,* vol. 13, no. 03, p. 1450027, 2014.

[32] 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.

[33] G. Cumming, Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis, Routledge, 2013.

[34] T. Elgamal and H. Mohamed , "Analysis of PCA algorithms in distributed environments," 13 5 2015. [Online]. Available: https://arxiv.org/abs/1503.05214. [Accessed 17 3 2015].

[35] A. Singh and V. Sarjolta, "MapReduce WordCount: Execution and Effects of Altering Parameters," *Int. J. Innov. Res. Comput. Commun. Eng.,* vol. 3, no. 10, p. 9330–9336, 2015.

[36] P.-N. Tan, M. Steinbach, A. Karpatne and V. Kumar, Introduction to data mining, New York: Pearson Education, 2018, pp. 327-414.

[37] U. M. L. Repository, "Thoracic Surgery Data Data Set," November 2013. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data.

[38] 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.

[39] M. Zięba, J. Tomczak, M. Lubicz and J. Świątek, "Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients," *Applied soft computing,* vol. 14, pp. 99-108, 2014.

[40] "Statlog (Heart) Data Set," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Statlog+(Heart).

[41] A. Asuncion and D. Newman, "UCI Machine Learning Repository," School of Information and Computer Science,University of California, Irvine, CA, 2007. [Online]. Available: http://archive.ics.uci.edu/ml/.

[42] W. H. Wolberg, "Breast Cancer Wisconsin (Original) Data Set," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original).

[43] P. Blake, R. Quinn and M. Oliver, "Peritoneal dialysis and the process of modality selection," *Peritoneal Dialysis International,* vol. 33, no. 3, pp. 233-241, 2013.

[44] M. Oliver, A. Garg, P. Blake, J. Johnson, M. Verrelli, J. Zecharias, S. Pandeya and R. Quinn, "Impact of contraindications, barriers to self-care and support on incident peritoneal dialysis utilization," *Nephrology Dialysis Transplantation,* vol. 25, no. 8, pp. 2737-2744, 2010.

[45] G. Michalopoulos, S. Subendran, Y. Yang, R. R. Quinn, M. J. Oliver, Z. Butt and A. Wong, "Interpretability of Machine Learning Models for Health Data - A Case Study," in *IMA2019:First International Workshop On Interpretability: Methodologies and Algorithms*, Adelaide, Australia, 2019.

[46] S. Ding, H. Zhu, W. Jia and C. Su, "A Survey On Feature Extraction for Pattern Recognition," *Artificial Intelligence Review,* vol. 37, no. 3, pp. 169-180, 2012.

[47] D. Powers, "Evaluation: From Precision, Recall and F-Measure To Roc, informedness, markedness and correlation," *J. Mach. Learn. Technol.,* vol. 2, pp. 37-63, 2011.

[48] D. K. Chiu, A. K. Wong and B. Cheung, "Information Discovery through Hierarchical Maximum Entropy Discretization and Synthesis," in *Knowledge Discovery in Databases*, 1991.

[49] R. Hodson, "Precision medicine," *Nature,* vol. 537, no. 7619, p. S49, 2016.

[50] Y. S. Koh and S. D. Ravana, "Unsupervised Rare Pattern Mining: A Survey," *ACM Trans. Knowl. Discov. Data,* vol. 10, no. 4, p. 1–29, 2016.

[51] 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.

[52] S. S. NIKAM, "A comparative study of classification techniques in data mining algorithms," *Oriental journal of computer science & technology,* vol. 8, no. 1, pp. 13-19, 2015.

[53] 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.

[54] Y. Sun, M. Kamel, A. K. Wong and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition,* vol. 40, no. 12, pp. 3358-3378, 2007.

[55] F. Cacheda, D. Fernandez, F. Novoa and V. Carneiro, "Eaarly Detection of Depression: Social Network Analysis and Random Forest Techniques," *Journal of Medical Internet Research,* vol. 21, no. 6, p. e12554, 2019.

[56] M. N. Parikh, H. Li and L. He, "Enhancing Diagnosis of Autism With Optimized Machine Learning Models and Personal Characteristic," *Frontiers in computational neuroscience,* vol. 13, no. 9, p. doi: 10.3389/fncom.2019.00009, 2019.

[57] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong , H. Shen and Y. Wang, "Artificial intelligence in healthcare: past, present and future," *Stroke Vasc Neurol,* vol. 2, no. 4, pp. 230-243, 2017.

[58] 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.

[59] S. Wang, "Mining Textural Features from Financial Reports for Corporate Bankruptcy Risk Assessment," M. Sc. Thesis, Systems Design Engineering, University of Waterloo, Waterloo, 2017.

[60] P. Keane and E. Topol , "With an eye to AI and autonomous diagnosis," *NPJ Digit. Med.,* vol. 1, no. 40, 2018.

[61] 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.