

Overview of Contrast Data Mining as a Field & Preview of an Upcoming Book

Guozhu Dong, Wright State University
James Bailey, University of Melbourne



Presented at International Workshop on Contrast Data Mining and Applications,
Part of the IEEE International Conference on Data Mining (ICDM), December 11 2011

Outline

- High level overview of contrasting
- General definitions and terminology
- Representative contrast pattern **mining algorithms**
- **Applications** of contrast mining for fundamental data mining tasks such as **classification and clustering**
- **Applications** of contrast mining in **bioinformatics, medicine, blog analysis, image analysis and subgroup mining**
- **Results** on **contrast based dataset similarity** measure, and on analyzing **item interaction in contrast patterns**
- Open research questions
- Overview of an **upcoming book**

High Level Overview of Contrasting

- Contrasting -- one of the most basic types of analysis
 - routinely used, by all types of people, perhaps subconsciously
 - to better understand the world around us
 - to better deal with the problems/challenges we face
- Contrasting involves the comparison of one set/kind/class of objects against another set/kind/class.
 - Aim: identify the differences b/w them, to provide useful insights on **how**, and perhaps also **why**, the **object classes differ**.

Contrasting can be employed in many situations/contexts

- two population groups – the young vs the elderly;
- two medical conditions -- the normal vs the diseased tissues of a cancer;
- two time periods -- performance of various groups/styles of stocks in 2009 vs in 2010;
- two spatial locations – states in US vs provinces in Canada;
- two DNA sequence locations – gene start sites vs non-gene start sites;
- analyzing holes and bumps in data;
- analyzing model shifts over time;
-

Traditional Approaches to Contrasting

- Before the age of computers, techniques for contrasting were often based on traditional statistical methods:
 - **comparing** the respective **means** of features of objects in the two sets
 - **comparing** the respective **distributions** of attribute values of objects in the two sets
- These approaches are limited -- difficult to use them to identify specific patterns that offer novel/actionable insights.
- In the last dozen years, significant progress on contrast data mining has been made.

General Definitions & Terminology (1)

- Given two datasets, $D1$ and $D2$, that one wishes to contrast, *contrast patterns are patterns that describe significant differences* between $D1$ and $D2$.
- A pattern X is considered as describing differences b/w the two datasets if *some statistics* (e.g., support or risk ratio) for X with respect to the datasets *are highly different*. E.G.
 - $|\text{supp1}(X) - \text{supp2}(X)|$ is large
 - $\text{supp1}(X) / \text{supp2}(X)$ is large
- We often refer to the dataset/class where a pattern P has the highest frequency as its *home dataset/class*.
- Contrast mining studied for many data types: transactions, vectors, sequences, graphs, images, texts, ...

General Definitions & Terminology (2)

- **Many names** have been used to describe contrast patterns, including *emerging patterns* [7], *contrast sets* [4], *group differences*, *patterns characterizing change*, *classification rules*, *discriminating patterns*, *conditional contrast* [39], ...
- Contrast patterns are often expressed as *conjunctions of simple conditions on attributes*; recent research has studied contrast patterns involving **more powerful** constructs:
 - *disjunctive emerging patterns* [23]
 - *fuzzy emerging patterns* [15]
 - *contrast inequalities* [11]
 - *contrast functions* [10]
 - *emerging cubes* [29]

Representative Contrast Pattern Mining Algorithms (1)

- A range of techniques/algorithms have been proposed.
- The algorithms often push pattern constraints (such as minimum/maximum frequency and minimum support difference/ratio) deep into the mining process.
- Data reduction based algorithms
 - Tree based algorithms [3]
 - Zero-suppressed binary decision diagram based algorithms [23]
- Lossless pattern space reduction algorithms
 - Border based algorithms [7]
 - Equivalence class based algorithms [19]

Some algorithms mine subset of CPs to be used as feature set for classifiers [35].
Some others mine other kinds of incomplete subsets [36,37].

Representative Contrast Pattern Mining Algorithms (2)

- For some data, e.g. microarray data for cancer, it may not be feasible or desirable to mine all contrast patterns
- Work in [28] presents a technique that mines desirable subsets of contrast patterns, using a **gene club based approach**
- A gene club for a given gene g is a set of genes that can differentiate between two different disease states and which are likely to interact with g with respect to the disease.
- This approach can mine some high quality (near optimal) contrast patterns involving each of the given genes, to offer insight on the role played by each of the genes in the disease.
- Very good performance – see next slide

EPs with high support for colon cancer found by Gene Club based Methods

A tissue is represented by ~2000 of features/genes

Colon Cancer EPs

{1+ 4- 112+ 113+} 100%
{1+ 4- 113+ 116+} 100%
{1+ 4- 113+ 221+} 100%
{1+ 4- 113+ 696+} 100%
{1+ 108- 112+ 113+} 100%
{1+ 108- 113+ 116+} 100%
{4- 108- 112+ 113+} 100%
{4- 109+ 113+ 700+} 100%
{4- 110+ 112+ 113+} 100%
{4- 112+ 113+ 700+} 100%
{4- 113+ 117+ 700+} 100%
{1+ 6+ 8- 700+} 97.5%

Colon Normal EPs

{12- 21- 35+ 40+ 137+ 254+} 100%
{12- 35+ 40+ 71- 137+ 254+} 100%
{20- 21- 35+ 137+ 254+} 100%
{20- 35+ 71- 137+ 254+} 100%
{5- 35+ 137+ 177+} 95.5%
{5- 35+ 137+ 254+} 95.5%
{5- 35+ 137+ 419-} 95.5%
{5- 137+ 177+ 309+} 95.5%
{5- 137+ 254+ 309+} 95.5%
{7- 21- 33+ 35+ 69+} 95.5%
{7- 21- 33+ 69+ 309+} 95.5%
{7- 21- 33+ 69+ 1261+} 95.5%

These EPs have 95%-100% support in one class but 0% support in the other class.

Minimal: Each proper subset occurs in both classes.

Only one of these EPs are be found from the top 70 genes (info gain order) .

None found using top 35 genes; highest home freq of EPs from top 35 genes was ~77%

1+: g1 is high

4-: g4 is low



Colon cancer dataset (Alon et al, 1999 (PNAS)): 40 cancer tissues, 22 normal tissues. 2000 genes



Applications of contrast mining for fundamental KDD tasks – classification (1)

- Since contrast patterns contain signals discriminating the classes, there have been **many studies** on how to use contrast patterns to build accurate classifiers.
- In general, three issues need to be addressed in order to build a contrast pattern based classification model:
 - contrast pattern mining, ← [covered by mining algs]
 - contrast pattern selection for use in classifier,
 - scoring strategy for the classification decision.
- Two main scoring approaches:
 - Deciding by **one** matching pattern for given sample t (e.g. CBA [25])
 - Decision by **all** matching patterns for t (aggregation) (e.g. CAEP [9])

Applications of contrast mining for fundamental KDD tasks – classification (2)

- CAEP: Classification by aggregating emerging patterns
 - Each matching EP for t contributes a discriminative vote/score for each class; class with highest aggregated score is the predicted class
- **Many variants** of CAEP have been proposed
 - iCAEP [34]: The predicted class for t is the one requiring the shortest EP based description length for t
 - DeEPs [20]: The predicted class for t is the class having the largest volume of data that match some matching EPs of t ; the EPs are mined on projected data that only contain items in t (lazy learner)
 - CPAR [34]: The predicted class of t is the class having highest average prediction among the best k contrast patterns that match t
- EPs with low support can be very useful here

Applications of contrast mining for fundamental KDD tasks – classification (3)

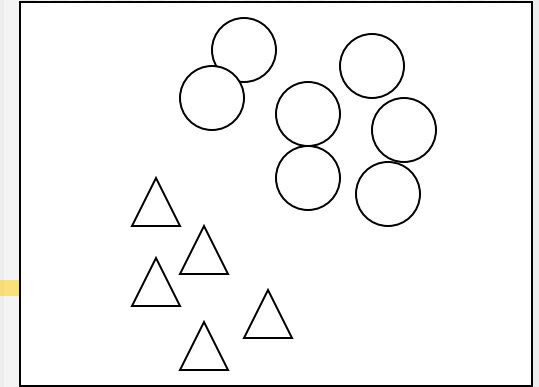
- Emerging patterns were also used to improve other classifiers
 - CAEP style EP based scores were used to assign weights to instances
 - Weight of a case t can indicate **confidence that t belongs to a class**; used to build weighted decision tree [2]
 - Weight of a case t can indicate **whether confidence on t 's class is ambiguous**; used to build weighted SVM [13]
- Emerging patterns were used to **add new training data instances** for rare class classification [1]
- Reference [5] proposed an **EP length statistics** based one-class classifier for outlier/intrusion detection
- Ref [16][17] proposed emerging pattern based methods to **classify images**.
 - Images are first partitioned into grids; then represented as transactions (of color/texture features) with occurrence counts.

Applications of contrast mining for fundamental KDD tasks -- clustering

- CPCQ clustering quality index [24]
- CPC clustering algorithm [12]
- They do not require distance functions.
- Distance functions are hard to define, since clustering is usually applied in explorative situations where little is known about the data.

The CPCQ

Clustering Quality Index



- A CP *characterizes* its home cluster and *distinguishes* its home cluster from other clusters.
- **CPCQ Rationale:**
 - A high-quality clustering has *many, diversified, high-quality* contrast patterns (CPs) for its clusters.
- CPCQ often recognizes expert-determined clusterings as superior to algorithm-determined clusterings (UCI data)

CP Quality and Diversity

- CP quality of a CP X : length ratio of closed pattern of X over minimal generator patterns of X .

- X 's EC = $\langle \{M_1, \dots, M_n\}, P \rangle$
- $\text{avg} \{ \text{length}(M_i) \mid i=1 \dots n \}$ is small \rightarrow easy to distinguish tuples in $\text{mat}(X)$ against other tuples
- $\text{length}(P)$ is large \rightarrow tuples in $\text{mat}(X)$ are highly coherent

EC of CPs: all CPs with the same set of matching data

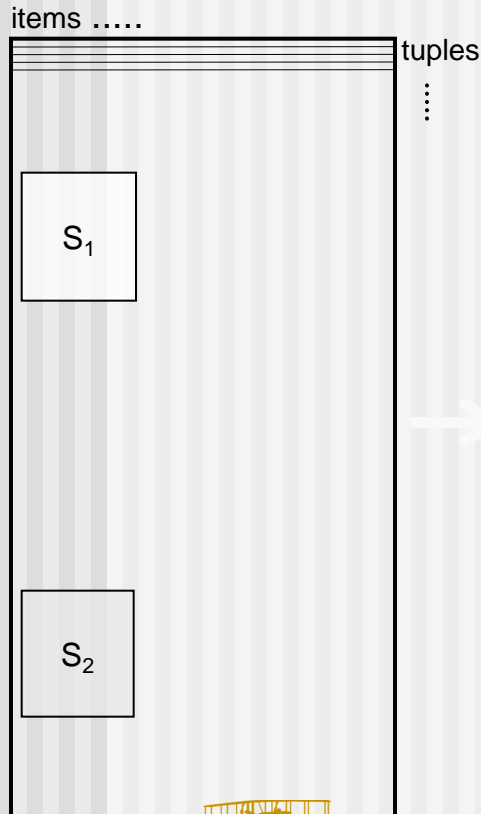
e.g. $\langle \{ab, ac, cd\}, abcdef \rangle$

- CP diversity: Many high quality CPs very dissimilar to each other

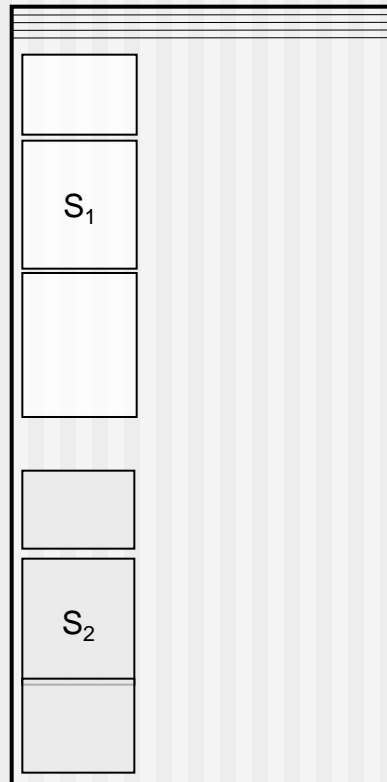
- High dissimilarity in both the items they contain and their $\text{mat}()$.
- There are many different ways to characterize tuples in the clusters

CPC Algorithm: Clustering to Maximize CPCQ

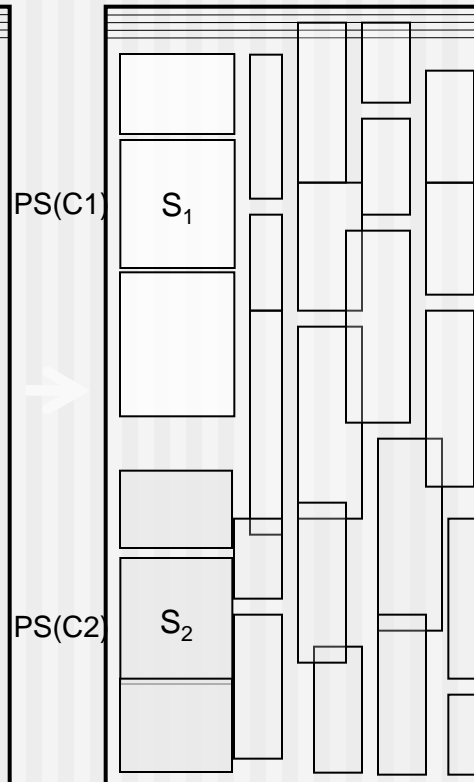
1. Select Seed CPs, using MPQ



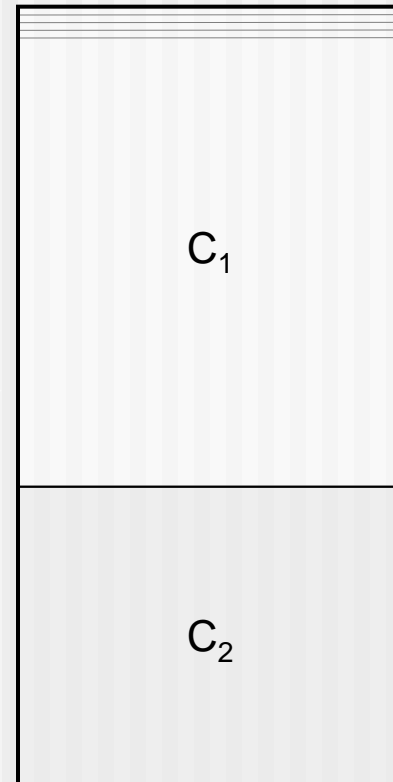
2. Assign Patterns to CP Group G_1 of Clusters, using MPQ



3. Assign Patterns as CPs of Clusters, using Tuple Overlap



4. Assign Tuples to Clusters, Using Tuple Overlap

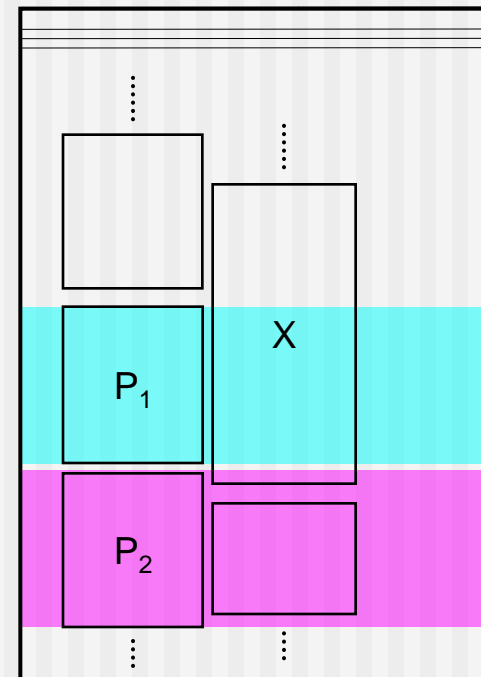
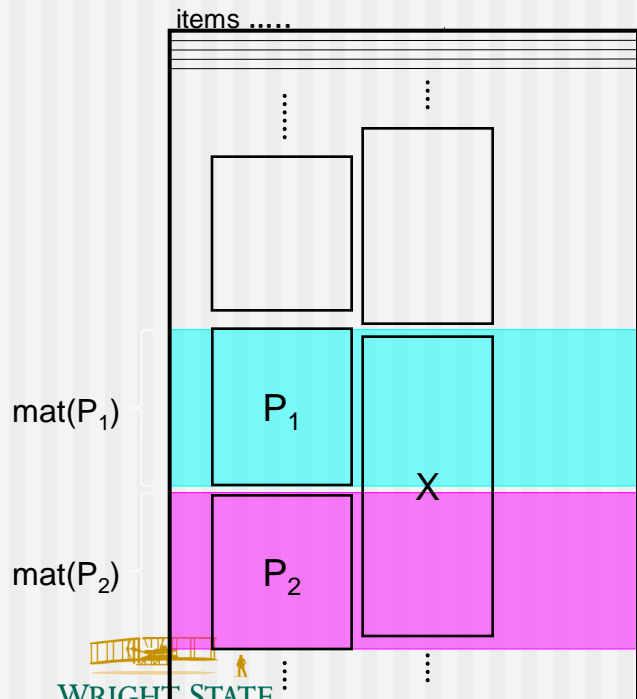


High quality mutual pattern vs low quality mutual pattern (MPQ)

- X is a **mutual pattern** of P1 & P2 if $\text{mat}(X)$ overlaps with both $\text{mat}(P_1)$ and $\text{mat}(P_2)$, but $\text{mat}(X)$ is not contained in either.

High-Quality Mutual Pattern X

Low-Quality Mutual Pattern X



tuples
⋮

Left:
• X becomes CP alongside P_1 & P_2 iff P_1 and P_2 are CPs of the same cluster.

Right:
• X may or may not be a CP in either case.

If many such X exist, then P1 and P2 should be in one cluster

Applications of contrast mining in bioinformatics, medicine, blog analysis, image analysis and subgroup mining (1)

- Microarray gene expression data based bioinformatics:
 - [22] used emerging patterns to characterize disease subtypes, used an emerging pattern based classifier to predict those subtypes.
 - [21] conjectured the possibility of using emerging patterns to design personalized treatment plans
 - [26] considered the use of contrast patterns to identify strong compound risk factors that have big risk differences.
 - [27] considered the use of transferability of discriminating genes (genes that occur in high quality emerging patterns) across microarray technology platforms to measure the concordance of microarray technology platforms.

Applications of contrast mining in bioinformatics, medicine, blog analysis, image analysis and subgroup mining (2)

- Blog community analysis.
 - [6] studied the use of contrast patterns as *distinct interest profiles* of communities of blogs.
 - It uses the CPC algorithm to form clusters of blogs based on their common distinct interest profiles, & uses a very small number of contrast patterns to characterize the discovered blog communities.
 - Useful for discovering and tracking blog communities based on their dynamic distinct interest profiles, instead of being based on the statically declared key words of interest of the blog authors.
- Image analysis: [16][17] -- discussed earlier
- Subgroup discovery and analysis: [30] examined the relationship between contrast patterns and subgroup mining and analysis. [38] studied mining differences between groups.

Blog Clustering/Description Experiments on data = Health U Music U Sports U Business

Blogs of four known categories were mixed, then clusters and cluster descriptions were generated by CPC

TABLE V. CLUSTER DESCRIPTIONS FOR CPC-GENERATED CLUSTERS, $k=4$, $\text{minS}=3\%$

Cluster	CPCQ-generated CP groups ($\text{minS}=3\%$) F-score: 0.77, CPCQ score: 0.325	
	DIP (CP group) 1	DIP (CP group) 2
1	{bodi, food}, {suffer, medic}	{symptom}, {health, fit}
2	{band, song}, {youtub, music}	{releas, song}
3	{team, game}	{season, team}
4	{busi, market}	{busi, monei}

Results on contrast based dataset similarity measure, and on analyzing item interaction in contrast patterns

- Work in [32] considered the use of cross dataset/class minimum coding length difference to define a similarity measure between datasets.
 - Here, encoding is done by using codes that represent patterns.
- **Comprehension and utility of contrast patterns** for domain experts is important. [14] analyzed the types of **item interactions** that may occur among items in contrast patterns, and categorized contrast patterns according to four types of item interaction
 - driver-passenger,
 - coherent,
 - independent additive,
 - synergistic beyond independent additive.

Open research questions

- Many challenges still remain and there is great potential for exciting research. Some open research questions for this field include:
 - *How does one assess the quality of contrast patterns, particularly for cases where the underlying datasets are of a complex type, such as a graph?*
 - *How can one incorporate domain knowledge to guide the discovery of contrast patterns?*
 - How can one use domain knowledge to understand the semantics of the mined contrast patterns, such as causation effects?
 - *Is it feasible and desirable to discover highly expressive contrast patterns, such as patterns defined by first order logic formulae?*
 - What other applications can benefit from contrast mining? How?

Overview of upcoming book (to appear in Spring 2012): ~ 20 chaps

- Mining algorithms (basic algorithms, incremental algorithms, mining subset of contrast patterns as features for classification, more expressive contrast pattern mining) – 6 chapters
- Contrast pattern based classification (basics chapter, image classification, enhancing classification, one-class classification) – 4 chapters
- Contrast pattern based clustering – 1 chapter
- Emerging cubes – 1 chapter
- Contrast patterns and microarray data analysis / bioinformatics -- 3 chapters
- Emerging patterns for chemoinformatics – 1 chapter
- Contrast patterns for geospatial applications – 1 chapter
- Emerging patterns for activity recognition – 1 chapter
- Emerging pattern and rough sets -- 1 chapter
- Other topics -- 1 chapter

www.cs.wright.edu/~gdong

ww2.cs.mu.oz.au/~jbailey/



Thank you !!!

References

- [1] Hamad Alhammady and Kotagiri Ramamohanarao. Using emerging patterns and decision trees in rare-class classification. In *IEEE International Conference on Data Mining (ICDM)*, pages 315–318, 2004.
- [2] Hamad Alhammady and Kotagiri Ramamohanarao. Using emerging patterns to construct weighted decision trees. *IEEE Trans. Knowl. Data Eng.*, 18(7):865–876, 2006.
- [3] James Bailey and Thomas Manoukian and Kotagiri Ramamohanarao. Fast Algorithms for Mining Emerging Patterns. In *European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 39-50, 2002.
- [4] Stephen D. Bay and Michael J. Pazzani. Detecting change in categorical data: Mining contrast sets. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 302–306, 1999.
- [5] Lijun Chen and Guozhu Dong. Masquerader detection using OCLEP: One class classification using length statistics of emerging patterns. In *International Workshop on Information Processing over Evolving Networks (WINPEN)*, 2006.
- [6] Guozhu Dong and Neil Fore. Discovering dynamic logical blog communities based on their distinct interest profiles. In *The First International Conference on Social Eco-Informatics (SOTICS 2011)*, 2011.
- [7] Guozhu Dong and Jinyan Li. Efficient mining of emerging patterns: Discovering trends and differences. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 43–52, 1999.
- [8] Wei Ding, Tomasz F. Stepinski, and Josue Salazar. Discovery of geospatial discriminating patterns from remote sensing datasets. In *SIAM International Conference on Data Mining (SDM)*, pages 425–436, 2009.
- [9] Guozhu Dong, Xiuzhen Zhang, Limsoon Wong, and Jinyan Li. CAEP: Classification by aggregating emerging patterns. In *Discovery Science*, pages 30–42, 1999.

References

- [10] Lei Duan, Changjie Tang, Liang Tang, Tianqing Zhang, and Jie Zuo. Mining class contrast functions by gene expression programming. In *International Conference on Advanced Data Mining and Applications (ADMA)*, pages 116–127, 2009.
- [11] Lei Duan, Jie Zuo, Tianqing Zhang, Jing Peng, and Jie Gong. Mining contrast inequalities in numeric dataset. In *International Conference on Web-Age Information Management (WAIM)*, pages 194–205, 2010.
- [12] Neil Fore and Guozhu Dong. CPC: A contrast pattern based clustering algorithm requiring no distance function. Technical report, Department of Computer Science and Engineering, Wright State University, 2011.
- [13] Hongjian Fan and Kotagiri Ramamohanarao. A weighting scheme based on emerging patterns for weighted support vector machines. In *IEEE International Conference on Granular Computing*, pages 435–440, 2005.
- [14] Gang Fang, Wen Wang, Benjamin Oatley, Brian Van Ness, Michael Steinbach, and Vipin Kumar. Characterizing discriminative patterns. *Computing Research Repository*, abs/1102.4, 2011.
- [15] Milton García-Borroto, José Francisco Martínez Trinidad, Jesús Ariel Carrasco-Ochoa. Fuzzy emerging patterns for classifying hard domains. *Knowledge and Information Systems*, 28(2):473489, 2011.
- [16] Lukasz Kobylinski and Krzysztof Walczak. Jumping emerging patterns with occurrence count in image classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 904–909, 2008.
- [17] Lukasz Kobylinski and Krzysztof Walczak. Spatial emerging patterns for scene classification. In *International Conference on Artificial Intelligence and Soft Computing*, pages 515–522, 2010.
- [18] Jinyan Li, Guozhu Dong, and Kotagiri Ramamohanarao. Making use of the most expressive jumping emerging patterns for classification. *Knowl. Inf. Syst.*, 3(2):131–145, 2001.

References

- [19] Jinyan Li and Guimei Liu and Limsoon Wong. Mining statistically important equivalence classes and delta-discriminative emerging patterns. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 430-439, 2007.
- [20] Jinyan Li, Guozhu Dong, Kotagiri Ramamohanarao, and Limsoon Wong. DeEPs: A new instance-based lazy discovery and classification system. *Machine Learning*, 54(2):99–124, 2004.
- [21] Jinyan Li and Limsoon Wong. Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. *Bioinformatics*, 18(10):1406–1407, 2002.
- [22] Jinyan Li, Huiqing Liu, James R. Downing, Allen Eng-Juh Yeoh, and Limsoon Wong. Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. *Bioinformatics*, 19(1):71–78, 2003.
- [23] Elsa Loekito and James Bailey. Fast Mining of High Dimensional Expressive Contrast Patterns Using Zero-suppressed Binary Decision Diagrams. Proceedings of The Twelfth ACM SIGKDD International Conference On Knowledge Discovery and Data Mining (KDD 2006). Pages 307-316, 2006.
- [24] Qingbao Liu and Guozhu Dong. A contrast pattern based clustering quality index for categorical data. In *IEEE International Conference on Data Mining (ICDM)*, pages 860–865, 2009.
- [25] Bing Liu, Wynne Hsu, and Yiming Ma. Integrating classification and association rule mining. In *KDD*, pages 80–86, 1998.
- [26] Jinyan Li and Qiang Yang. Strong compound-risk factors: Efficient discovery through emerging patterns and contrast sets. *IEEE Transactions on Information Technology in Biomedicine*, 11(5):544–552, 2007.
- [27] Shihong Mao, Charles Wang, and Guozhu Dong. Evaluation of inter-laboratory and cross-platform concordance of dna microarrays through discriminating genes and classifier transferability. *J. Bioinformatics and Computational Biology*, 7(1):157–173, 2009.

References

- [28] Shihong Mao and Guozhu Dong. Discovery of Highly Differentiative Gene Groups from Microarray Gene Expression Data Using the Gene Club Approach. In *J. Bioinformatics and Computational Biology*, 3(6), 1263–1280, 2005.
- [29] Sébastien Nedjar, Rosine Cicchetti, and Lotfi Lakhal. Extracting semantics in OLAP databases using emerging cubes. *Information Sciences*, 2011.
- [30] Petra Kralj Novak, Nada Lavrac, and Geoffrey I. Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10:377–403, 2009.
- [31] Pawel Terlecki. On the relation between jumping emerging patterns and rough set theory with application to data classification. *Transactions on Rough Sets XII*, 12:236–338, 2010.
- [32] Jilles Vreeken, Matthijs van Leeuwen, and Arno Siebes. Characterising the difference. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 765–774, 2007.
- [33] Xiaoxin Yin and Jiawei Han. CPAR: Classification based on predictive association rules. In *SIAM International Conference on Data Mining (SDM)*, 2003.
- [34] Xiuzhen Zhang, Guozhu Dong, and Kotagiri Ramamohanarao. Information-based classification by aggregating emerging patterns. In *Intelligent Data Engineering and Automated Learning (IDEAL)*, pages 48–53, 2000.
- [35] Hong Cheng, Xifeng Yan, Jiawei Han, and Philip S. Yu. Direct discriminative pattern mining for effective classification. In *IEEE International Conference on Data Engineering*, pages 169–178, 2008.
- [36] Stephen D. Bay and Michael J. Pazzani. Detecting change in categorical data: Mining contrast sets. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 302–306, 1999.

References

- [37] Geoffrey I. Webb. Discovering significant patterns. *Machine Learning*, 68(1):1-33, 2007.
- [38] Geoffrey I. Webb, Shane M. Butler, and Douglas A. Newlands. On detecting differences between groups. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 256{265, 2003.
- [39] Guozhu Dong, Jinyan Li, Guimei Liu, and Limsoon Wong. Mining Conditional Contrast Patterns. In *Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction*. Edited by Yanchang Zhao and Chengqi Zhang and Longbing Cao. IGI Global, 2009.

Example EP in microarray data for cancer

Normal Tissues

Cancer Tissues

binned
(intervalized)
data

genes →

tissues ↘

g1	g2	g3	g4
L	H	L	H
L	H	L	L
H	L	L	H
L	H	H	L

g1	g2	g3	g4
H	H	L	H
L	H	H	H
L	L	L	H
H	H	H	L

EP example: $X = \{g1=L, g2=H, g3=L\}$; $\text{suppN}(X) = 50\%$, $\text{suppC}(X) = 0$

Or $X = \{1-, 2+, 3-\}$