SciCSM: Novel Contrast Set Mining over Scientific Datasets Using Bitmap Indices

Gangyi Zhu, Yi Wang, Gagan Agrawal

The Ohio State University

Outline

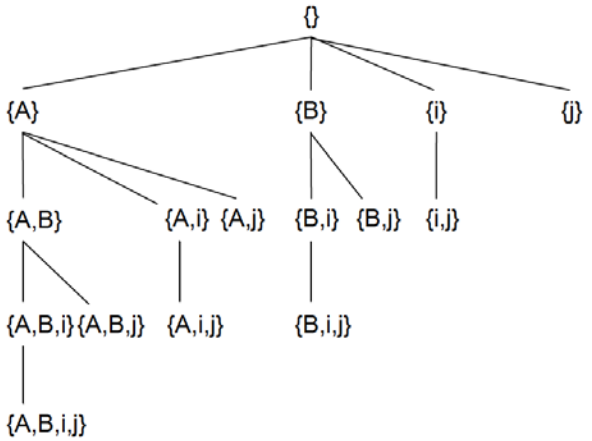
- Introduction
- SciCSM Algorithm
- Pruning and Combination
- Bitmap-Based Optimization
- Experiment Results
- Conclusion

Introduction

- Goal
  - Identify all the interesting contrast sets, which are described as conjunctions of attributes value pairs and have significant difference in different groups
- Example
  - Given two datasets of climate simulation for the same spatial area but different time periods, the key differences between two datasets, which can be taken as climate changes, is of great interest
- Challenges for Array-based Scientific Data
  - Both value-based and dimension-based attributes are involved
  - The attributes are mostly numeric for scientific datasets
  - Scientific array dataset of interest to us can be extremely large

SciCSM Algorithm

- Search Strategy
  - Discretize each attribute range into a number of bins
  - A set-enumeration tree is adopted as foundation of search strategy
  - All attributes are enumerated in a certain order, and every node represents a contrast set
- Exhaustive but Efficient Search
  - Level-wise(top-down)
  - Pruning + combination
  - Each node is visited only once or not visited at all (if pruned)



Pruning and Combination

- Pruning Strategy
  - **Minimum Absolute Quality**: sufficient quality
  - **Minimum Relative Quality**: sufficient quality improvement compared with its parent
- Combination Strategy
  - **Existence of Adjacency**: only adjacent nodes can be combined
  - **Combined Quality:** combined node should have sufficient quality
  - **Difference Homogeneity:** support/mean difference in two combining nodes should have the same trend (both positive or negative)
  - **Distribution Purity**: introduce a measure Continuous Weighted Entropy (CWE) to control the granularity of output contrast sets

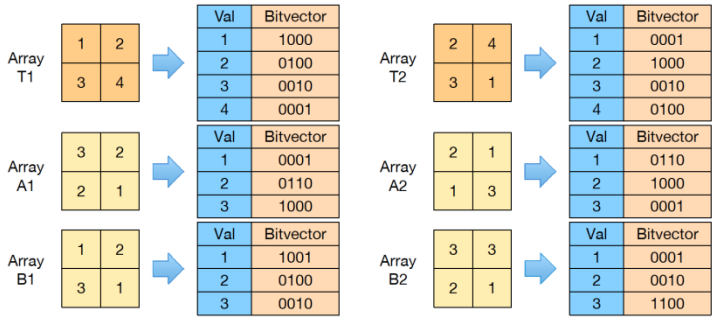
$$CWE = -\sum_{i=1}^{n} p(i) \times \log p(i) \times \frac{|v_i - M|}{v_n - v_1}$$

Entropy

Weight

Bitmap-Based Optimization

- Our algorithm entirely operates on bitmap indices
- Key Insights
  - Each attribute-range pair can be represented by a disjunction of bitvectors
  - Each contrast set can be represented by a conjunction of bitvectors
  - All the statistics like mean, CWE, and quality can be calculated entirely based on bitmap



(a) Conventional Indices

Index	Val	Bitvector1	Bitvector2
Bitvector for Dimension-based i	1	1100	1100
	2	0011	0011
Bitvector for Dimension-based j	1	1010	1010
	2	0101	0101
Bitvector for Value-based A	[1,2]	0111	1110
	[3,4]	1000	0001
Bitvector for Value-based B	[1,2]	1101	0011
	[3,4]	0010	1100

(b) Indices for Dimension-based Attributes and Value-based Attributes Bitvectors

A Bitmap Indexing Example in SciCSM

Experiment Results

- Hardware: A machine with 48 GB of main memory and Intel® Xeon(R) CPU X5650 2.67 GHz CPU

- Quality Evaluation

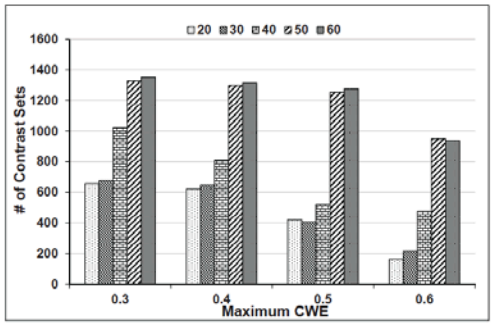
Table 1: Best 8 Contrast Sets Discovered by SciCSM on WOA13_1DEG with Maximum CWE of 0.50

Rank	Contrast Set	Quality	Mean Diff.	Sup. Diff.
1	$0.05 \leq$ AOU < 0.64	1.076	0.58%	7.00%
2	$14.47 \leq$ TEMP < 15.00 *AND* $35.25 \leq$ SALT < 35.37	1.058	5.78%	0.04%
3	$-2.62 \leq$ AOU < 0.05	1.056	1.04%	4.47%
4	$14.47 \leq$ TEMP < 15.00 *AND* $34.88 \leq$ SALT < 34.91	1.049	4.94%	0.00%
5	$14.47 \leq$ TEMP < 15.00 *AND* $35.90 \leq$ SALT < 41.04	1.039	3.75%	0.17%
6	$21.46 \leq$ TEMP < 37.74	1.034	0.94%	2.44%
7	$34.92 \leq$ SALT < 34.94 *AND* $-179.50 \leq$ lon < -120.50	1.033	3.26%	0.00%
8	$-1.92 \leq$ TEMP < 1.25	1.031	1.87%	1.23%

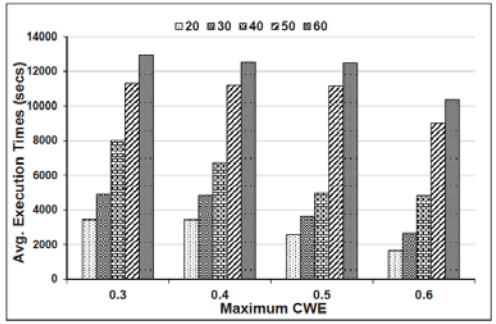
Table 2: Best 8 Contrast Sets Discovered by SciCSM on WOA13_1DEG with Maximum CWE of 0.75

Rank	Contrast Set	Quality	Mean Diff.	Sup. Diff.
1	$14.47 \leq$ TEMP < 15.00 *AND* $35.08 \leq$ SALT < 35.15	1.140	13.95%	0.02%
2	$14.47 \leq$ TEMP < 15.00 *AND* $35.50 \leq$ SALT < 35.62	1.089	8.92%	0.01%
3	$14.47 \leq$ TEMP < 15.00 *AND* $35.20 \leq$ SALT < 35.44	1.082	8.16%	0.07%
4	$0.05 \leq$ AOU < 0.64	1.076	0.58%	7.00%
5	$-2.62 \leq$ AOU < 0.05	1.056	1.04%	4.47%
6	$14.47 \leq$ TEMP < 15.00 *AND* $35.78 \leq$ SALT < 41.04	1.040	3.83%	0.19%
7	$-1.92 \leq$ TEMP < 2.14	1.037	2.31%	1.40%
8	$14.47 \leq$ TEMP < 15.00 *AND* $34.85 \leq$ SALT < 34.94	1.037	3.67%	0.01%

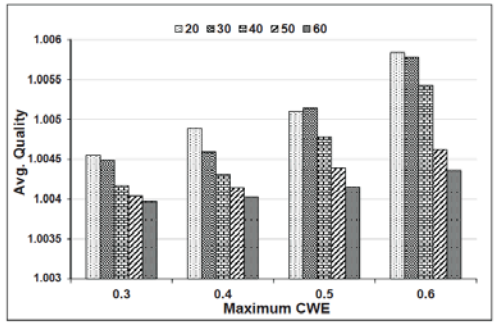
- Performance Evaluation



(a) # of Contrast Sets on NARCCAP



(b) Avg. Exe Times on NARCCAP



(c) Avg. Quality on NARCCAP

Conclusion

- Functionality
  - Capable of handling arrays comprising value-based attributes and dimension-based attributes
  - Can effectively process numeric attributes
- Bitmap Acceleration
  - Compact input leads to I/O and storage advantage
  - Fast bitwise operations support highly efficient computation
  - Provide ability of processing larger datasets
- Efficiency and Effectiveness
  - Both high efficiency and effectiveness are demonstrated by extensive experiments on multiple real-life datasets