Institute for
Data Science & Informatics
University of Missouri

Electrical Engineering
& Computer Science
University of Missouri

INTERDISCIPLINARY DATA ANALYTICS AND SEARCH

Exploratory Data Mining for Subgroup Cohort Discoveries and Prioritization

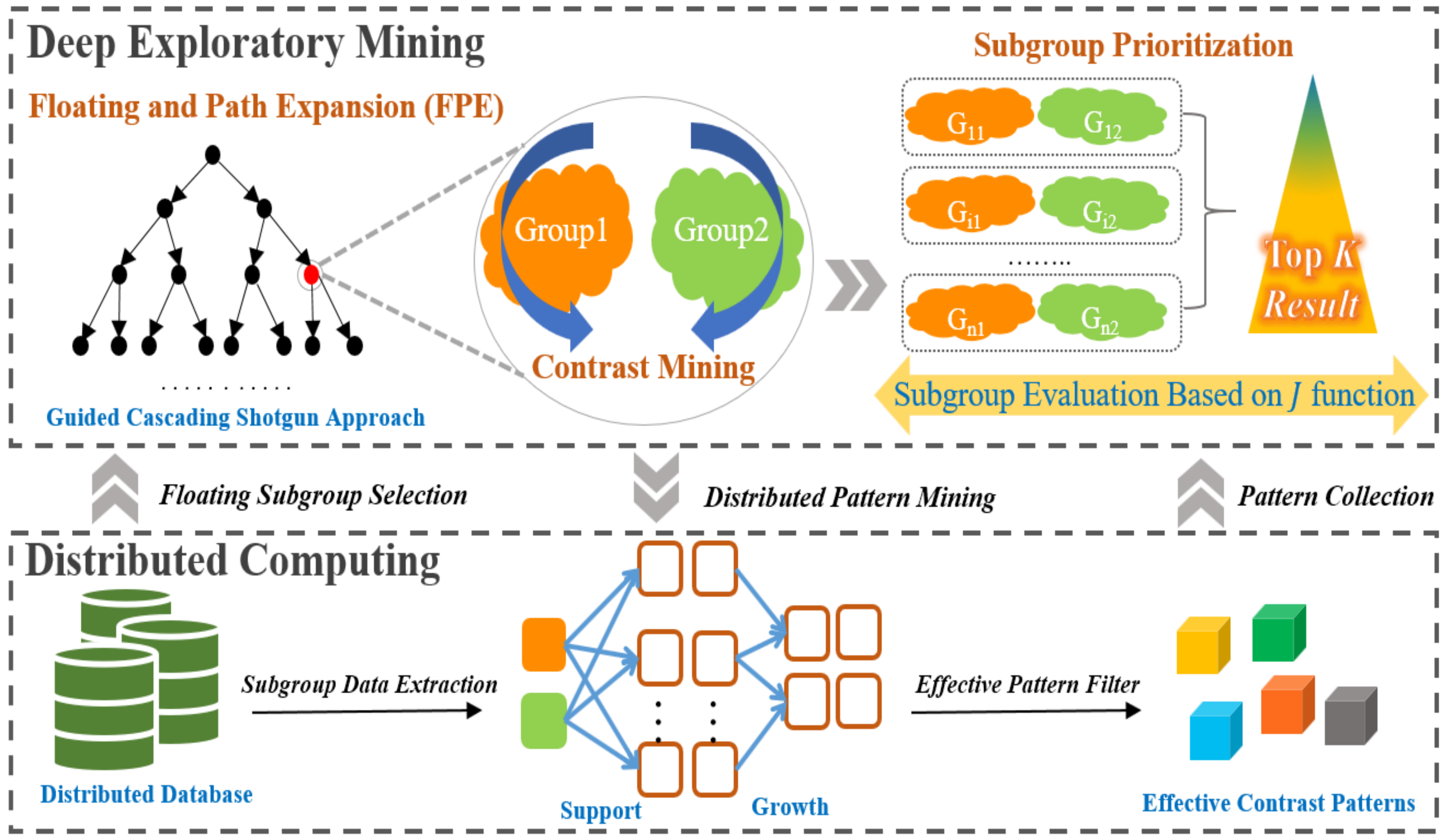
Danlu Liu, William Baskett, Chi-Ren Shyu

Introduction

Finding small homogeneous subgroup cohorts in large heterogeneous populations is a critical process for hypothesis development in biomedical research. Concurrent computational approaches are still lacking in robust answers to the question "what hypotheses are likely to be novel and to produce clinically relevant results with well thought-out study designs?" The goal of exploratory data mining for cohort discovery is to provide a robust data-driven framework to tailor potential interventions for precision health automatically.

Method

Our floating and path expansion approach utilizes a less greedy and more computational feasible floating selection process to select potential subgroups based on their qualities, and the expansion process controls the ratio of the number of subgroups for later floating iterations. Then, we use a distributed computing framework to identify contrast patterns of features which differentiate groups by exploring patterns that have an imbalanced prevalence between the groups. Finally, we prioritize the candidate subgroup pairs by J-value, an index for evaluating the aggregated contributions of the extracted contrast patterns within each pair of subgroups based on the number of contrast patterns and the significance of those patterns.



What hypotheses are likely to be novel and to produce clinically relevant results with well thought-out study designs?

Take a picture to
download the full paper

D. Liu, W. Baskett, D. Beversdorf and C. Shyu, "Exploratory Data Mining for Subgroup Cohort Discoveries and Prioritization," in *IEEE Journal of Biomedical and Health Informatics*. doi: 10.1109/JBHI.2019.2939149

Results

We used the Simons Foundation Autism Research Initiative (SFARI) Simon's Simplex Collection (SSC) for autism cohort discoveries. The data contains 2591 families with exactly one child diagnosed with autism (proband) while the parents and siblings are unaffected. By performing the deep exploratory data mining method with a 20% expanding factor, we discovered 142 contrast subgroups. From all discovered genes or gene combinations in the top 20 subgroup cohorts, 11.57% of 415 relevant genes are in AutDB, nearly 20.72% were identified through the PubMed search, and the remaining genes were considered novel.

TABLE I
RATED CONTRAST SUBGROUPS AND RATIO OF PUBLISHED SIGNIFICANT GENES

Subgroup 1 [a]			Subgroup 2		No. of Discovered Genes	No. of Discovered Genes also in AutDB [b]	No. of PubMed Articles
Population Variable(s)	Cohort Size		Population Variable(s)	Cohort Size	Number	Number	Number
Low SSC Full Scale IQ	459	vs	High SSC Full Scale IQ	373	5	1	2242
Normal to Speak Sentences	346	vs	Late to Speak Sentences	304	16	3	5130
Mid RBS-R Overall Score **AND** Low CBCL6 Social Score	202	vs	Low RBS-R Overall Score **AND** Low CBCL6 Social Score	77	44	6	898
Low ABC III Stereotypy Scale **AND** Late to Use Words	171	vs	High ABC III Stereotypy Scale **AND** Late to Use Words	159	18	2	452
Mid Vineland II Daily Living **AND** High Height Z Score **AND** High ADIR C Total	253	vs	High Vineland II Daily Living **AND** High Height Z Score **AND** High ADIR C Total	54	22	4	0
Mid CBCL6 Rule Breaking Score **AND** Low CBCL6 Activities Score **AND** High SRS-P Total Score	228	vs	High CBCL6 Rule Breaking Score **AND** Low CBCL6 Activities Score **AND** High SRS-P Total Score	59	25	4	0

[a] SSC Full Scale IQ=Simons Simplex Complex Full Scale IQ, RBS-R=Repetitive Behaviors Scale-Revised, CBCL6=Child Behavior Checklist for ages 6-18, ABC III=Aberrant Behavior Checklist-Stereotype Scale, Vineland II Daily Living=Vineland Adaptive Behavior Scales-Second Edition in Daily Living domain, ADIR C Total=Autism Diagnostic Interview-Revised (ADI-R)-Restricted, Repetitive, and Stereotyped Patterns of Behavior total score, SRS-P=Social Responsiveness Scale – Parent Report.

[b] Details about significant genes are in the Supplement 1.

Conclusion

The success of identifying optimal subpopulations is expected to result in much more promising findings for tailoring treatment than simply looking at the population as a whole for precision health research. This framework will provide the broad biomedical research community with a means to develop strategies to identify homogeneous subgroups within heterogeneous populations prior to conducting costly bench experiments or clinical trials. It has the potential to enable targeted treatments to improve outcomes, reduce costs, and minimize morbidity associated with misdirected interventions.

University of Missouri