Query-constraint-based Association Rule Mining from Diverse Clinical Datasets in the National Sleep Research Resource

Rashmie Abeysinghe*, Licong Cui*†
*Department of Computer Science
†Institute for Biomedical Informatics
University of Kentucky, Lexington, Kentucky, USA

***Abstract*—Secondary use of biomedical data has gained much attention recently to facilitate rapid knowledge discovery in biomedicine. Association Rule Mining (ARM) has been a popular technique for biomedical researchers to perform exploratory data analysis and discover potential relationships among variables in biomedical datasets. However, ARM of a high-dimensional biomedical dataset may produce a large number of rules that may not be interesting. In this paper, we introduce a query-constraint-based ARM (QARM) approach for exploratory analysis of diverse clinical datasets integrated in the National Sleep Research Resource (NSRR), which enables the rule mining on a subset of data containing items of interest based on a query constraint. In addition, biomedical datasets always contain semantically similar variables, thus we performed similar-variable-merging so that rules with simlar variables are not obtained. Applying QARM on five datasets from NSRR obtained a total of 6,921 rules with a minimum confidence of 60% (using top 50 rules for each query constraint).**

***Index Terms*—Query-constraint-based Association Rule Mining; National Sleep Research Resource; Exploratory Data Analysis**

I. Introduction

The ever expanding content of large clinical and biomedical datasets provides researchers with significant opportunities for data analysis and knowledge discovery in biomedicine [1]. One such data repository is the National Sleep Research Resource (NSRR), which is funded by the National Heart, Lung, and Blood Institute (NHLBI). NSRR has been designed to provide big data resources to the sleep research community [2]. Repositories like NSRR, if used properly, could aid in informed decision making and improve patient safety [3]. From a research perspective, they could be used in knowledge discovery to facilitate rapid generation or testing of hypotheses.

Association Rule Mining (ARM) is an exploratory data mining technique that has shown great potential in the biomedical domain for knowledge discovery. Given some predefined interestingness parameters, it is used extensively to find associations among variables. One of the potential issues with ARM in biomedical research is the number of rules that it yields could be humongous for large biomedical datasets. Most of these rules obtained could be medically irrelevant or uninteresting. This is often the case if the basic ARM is performed directly on a biomedical dataset. Another potential challenge in mining biomedical datasets is the existence of semantically similar variables. Rules containing such similar variables are of less interest because these variables capture similar or same characteristics. Therefore, it is often needed to apply certain techniques to filter out those irrelevant or uninteresting rules.

In this paper, we present QARM, a query-constraint-based ARM method to generate association rules based on a subset of data that satisfies certain query constraints or criteria. For example, if the criteria was "having had a stroke", then the generation of association rules would be only based on the patients who have had a stroke; in this way, the rules obtained would be more relevant to the criteria of interest. Such query-criteria-based ARM empowers biomedical researchers to perform exploratory data analysis in large biomedical data repositories and generate or test potential hypotheses.

The remainder of this paper is structured as follows. Section II provides some background information on NSRR, ARM and the related work on association rule mining in biomedical domain. Section III describes the query-constraint-based method that was undertaken. The results can be found in Section IV. Section V contains a discussion about the results and future directions. Section VI concludes this paper.

II. Background

A. National Sleep Research Resource (NSRR)

NSRR, launched in 2014, provides access free of charge in a web-based portal to large collections of de-identified physiological signals and clinical data elements (or variables) collected in well-characterized cohorts and clinical trials to support research on risk factors and outcomes of sleep disorders [4]. Each de-identified patient record of NSRR contains clinical data elements including demographic information, anthropometric parameters, physiologic measurements, medical history, medications, sleep symptoms, and other symptoms [2].

In this work, we use five datasets from NSRR: Cleveland Family Study (CFS), Childhood Adenotonsillectomy Trial (CHAT), Hispanic Community Health Study/Study of Latinos (HCHS/SOL), Heart Biomarker Evaluation in Apnea Treatment (HeartBEAT), and Sleep Heart Health Study (SHHS).

This work was supported by the National Science Foundation through grant 1657306. Correspondence: licong.cui@uky.edu

For each dataset in NSRR, the clinical data as well as the data dictionary (which contains metadata) are stored in comma-separated values (CSV) files. The common data elements across different NSRR datasets are maintained in a Canonical Data Dictionary (CDD), and mappings are provided between the CDD elements and the data elements in each individual dataset.

B. Association Rule Mining (ARM)

Agrawal et al. [5] originally developed ARM with the purpose of market-based analysis where patterns like $X \rightarrow Y$ (e.g., $\{Bread\} \rightarrow \{Butter\}$) are of importance. The intuitive meaning of this is that transactions which contain X tend to contain Y as well. A transaction corresponds to the products that a customer bought in a single visit to a store. Individual products are called items in the transaction. This approach is very flexible and general enough to be applied in many areas.

There are two parameters which correspond to each mined association rule: *Support* and *Confidence*.

- Support: $support(X \rightarrow Y) = Prob(X \cup Y) = support(X \cup Y)$.
- Confidence: $confidence(X \rightarrow Y) = Prob(Y|X) = support(X \cup Y)/support(X)$.

Rules that satisfy the user-specified minimum support (*minsup*) and minimum confidence (*minconf*) thresholds are called strong association rules.

There are many algorithms introduced for ARM [6], [7]. In this work we leverage the top-k non redundant association rules algorithm [8].

C. Top-k Non-Redundant (TNR) Association Rule Mining Algorithm

Fournier-viger et al. [8] introduced the top-k algorithm to address the difficulty in selecting suitable values to parameters *minsupp* and *minconf*. In our work, especially since each query constraint corresponds to a subset of patient records, fine-tuning *minsupp* and *minconf* parameters so that, any given query results in a satisfactory number of rules is a difficult task.

Fournier-viger et al. [9] later introduced the TNR algorithm to address the redundancy issues existing in the original top-k algorithm. The TNR algorithm takes k (the number of association rules to be found), *minconf* and Δ (exactness improving parameter) as parameters, and approximates top-k rules with the top support having a confidence above the *minconf* threshold.

D. Related Work

ARM has been applied to biomedical domains to facilitate knowledge discovery and disease prediction. For example, Hristovski et al. [10] have presented an interactive biomedical discovery support system which is capable of discovering new, potentially meaningful relations between a given starting concept of interest and other concepts. Hu et al. [11] have introduced a semantic-based ARM method to discover hidden connections among biomedical concepts from disjoint biomedical literature sets. Ordonez et al. [12] have introduced a method to predict heart diseases by ARM. They have analyzed the idea of discovering constrained association rules in medical records. Ordonez et al. [13] have introduced an ARM method that uses search constraints to reduce the number of rules.

III. Methods

In this work, we introduce QARM, a query-constraint-based ARM method for exploratory analysis of biomedical datasets. First we perform a series of steps for data pre-processing, including variable selection, variable merging, combining multiple-visit data, and query-constraint-based data transformation. Then we apply QARM using the top-k non-redundant ARM algorithm to the five datasets in NSRR.

A. Variable Selection

Each variable in NSRR datasets has a type (e.g., categorical, numerical). The possible values for a categorical variable fall into several predefined distinct classes. Therefore, categorical variables in NSRR datasets have domains defining the distinct classes into which their values could fall.

In this work, we mainly focused on categorial variables with domains of the yes/no type for simplicity. In addition, we selected variables regarding patients' medical history, medications, sleep symptoms, and other symptoms.

Based on the above variable selection criteria, we obtained a set of selected variables from the Canonical Data Dictionary (called *canonical variables*), as well as the study-specific variables which are mapped to the canonical variables for each individual dataset (called *dataset variables*). It is worth noting that one canonical variable may map to multiple dataset variables.

B. Variable Merging

Since certain variables in a dataset may capture similar information, association rules obtained including such similar variables would be of less interest. For example, both variables *prev_hx_stroke (Previous history of stroke)* and *stroke15 (MD Reported Stroke)* in SHHS mapping to the canonical variable *strokehist (Stroke - history)* capture the information about whether a patient has had a stroke. Occurrences of such variables together in a rule might make it uninteresting.

Therefore, to avoid obtaining association rules with such similar variables, we performed QARM with merging of such variables so that whenever a patient exhibits a "yes" to at least one of the similar variables, then the value of the merged variable will also be "yes". Here, the dataset variables mapping to the same canonical variable are considered similar, and hence merged. We refer to this method as the "merged method".

C. Combining Multiple-visit Data

In NSRR, a dataset may contain patient data with multiple visits. For instance, the datasets CHAT, HeartBEAT and SHHS contain data collected in two patient visits. These multiple

visits of a dataset were combined into one as a preprocessing step before QARM was performed. Since multiple visits may contain data collected for the same variable, we performed the combination in the following way: for the same patient, if the value of the variable appear as "yes" in at least one of the visits, then the combined result will be "yes"; otherwise, the combined result will be "no".

D. Query-constraint-based Data Transformation

Given a query constraint, the clinical data of patients satisfying the query criteria needs to be transformed to a suitable format before being fed into the TNR algorithm. In clinical datasets like NSRR, the possible values of a patient variable with the domain of yes/no type may be "yes", "no", or "unknown" (or "NA"). While "no" and "unknown" are important for capturing more precise information of patients, they may not be useful for generating association rules. If such "no" values were used for generating association rules (denoting the characteristics patients do not have), then it would have produced a lot of uninteresting and irrelevant rules also making the ARM process slow. Therefore, in this work, we only consider variables with "yes" value for each patient record satisfying the query criteria.

E. QARM using TNR Algorithm

Given a query constraint, QARM using TNR algorithm was applied to the patient data satisfying the query constraint after data transformation, with $k = 50$, $minconf = 60\%$ and $\Delta = 10$. We set a lower-bound of 20 to the number of patient records exhibiting this query constraint characteristic as a condition for the applicability of QARM so that a sufficient number of patient records will be considered.

Note that the support and the confidence of the rules obtained are based on a sub-dataset of patients satisfying the query constraint, not the entire dataset. In addition, the query constraint itself was not used for performing QARM since it is satisfied by each patient record in the sub-dataset.

IV. Results

A total of 71 canonical variables were obtained after the variable selection process. Since each canonical variable can serve as a query constraint, we interchangeably use terms "canonical variable" and "query constraint" in the followings. Table I shows the numbers of canonical variables identified in each of the five datasets, the numbers of mapped dataset variables corresponding to the canonical variables, and the numbers of association rules obtained within each dataset. It can be seen that SHHS covered the most number of canonical variables.

A total of 6,921 association rules were obtained by applying QARM within each of the five datasets, using top $k = 50$ rules with a *minconf* threshold of 60%.

Table II gives ten examples of association rules for the query constraint *cancerhist (Cancer-history)* in the HeartBEAT dataset.

TABLE I
NUMBER OF CANONICAL VARIABLES USED IN EACH DATASET, NUMBER OF DATASET VARIABLES THE CANONICAL VARIABLES MAP TO AND THE NUMBER OF ASSOCIATION RULES OBTAINED.

Dataset	No. of canonical variables	No. of mapped dataset variables	No. of association rules
CFS	40	113	2,000
CHAT	5	20	221
HCHS/SOL	31	75	1,550
HeartBEAT	13	31	650
SHHS	50	138	2,500

TABLE II
EXAMPLES OF ASSOCIATION RULES FOR THE QUERY CONSTRAINT"*cancerhist (Cancer-history)*" IN THE HEARTBEAT DATASET.

Antecedent	Consequent
Habitual Snoring	Hypertension-history
Hypertension-history	Habitual Snoring
Hypercholesterolemia-history	Hypertension-history
Hypertension-history	Hypercholesterolemia-history
Hypercholesterolemia-history, Habitual Snoring	Hypertension-history
Hypercholesterolemia-history, Hypertension-history	Habitual Snoring
Hypercholesterolemia-history	Habitual Snoring
Hypercholesterolemia-history	Habitual Snoring, Hypertension-history
Habitual Snoring, Hypertension-history	Hypercholesterolemia-history
Habitual Snoring	Hypercholesterolemia-history, Hypertension-history

In QARM, the same query constraint may generate different association rules for disparate datasets. Table III shows the numbers of common and distinct rules among five datasets for 10 query constraints, respectively.

TABLE III
NUMBERS OF COMMON AND DISTINCT RULES OBTAINED FOR DIFFERENT DATASETS FOR 10 QUERY CONSTRAINTS.

Description	Query constraint	Datasets	No. of common rules	No. of distinct rules for each dataset
Anxiety disorder	anixietyhist	CFS, HEARTBEAT	8	42
Hypercholesterolemia-history	cholesthist	CFS, HEARTBEAT	7	43
Chronic cough	chroniccough	CFS, HCHS	5	45
Depression	depresshist	CFS, HEARTBEAT	5	45
Angina pectoris	anginahist	SHHS, CFS	5	45
Cancer-history	cancerhist	CFS, HEARTBEAT	4	46
Histamine-2 Receptor Antagonist	h2blocker	SHHS, CFS	4	46
Cardiovascular disease - history	cvdishist	SHHS, CFS	3	47
Proton Pump Inhibitor	ppi	SHHS, CFS	3	47
Persistent wheezing	persistwheez	CFS, HCHS	3	47

V. Discussion

In this work, we investigate QARM, a query-constraint-based ARM method which is applied to five clinical datasets in NSRR. From the results, it can be seen that for the same query constraint, most rules obtained for an individual dataset are unique to that particular dataset. This may be due to the distinct patient characteristics captured in disparate datasets.

A. Distinction with Related Work

ARM has been widely applied to biomedical datasets for data-driven knowledge discovery. However, exploratory ARM based on a particular query constraint has been rarely investigated. QARM would allow researchers to perform exploratory analysis based on the data of interest by composing a specific query criteria to filter out irrelevant data.

B. Limitations and Future Work

In this work, we only considered categorical variables for the query-constraint-based ARM. It would be interesting to further investigate numerical variables, where numerical values can be categorized into some predefined ranges. It would also be interesting to incorporate variables with domains of other than *yes/no* type. In addition, we would like to generalize QARM, so that it allows multiple canonical variables as query constraints.

VI. Conclusion

In this paper, we applied QARM, a query-constraint-based association rule mining method, to five diverse clinical datasets in the National Sleep Resource Resource. QARM shows the potential to support exploratory analysis of large biomedical datasets by mining a subset of data satisfying a query constraint. It is also shown as a useful method to obtain interesting association rules from imbalanced datasets.

References

[1] X. Wang, M. R. Smith, and R. M. Rangayyan, "Mammographic information analysis through association-rule mining," in *Electrical and Computer Engineering, 2004. Canadian Conference on*, vol. 3. IEEE, 2004, pp. 1495–1498.

[2] D. A. Dean, A. L. Goldberger, R. Mueller, M. Kim, M. Rueschman, D. Mobley, S. S. Sahoo, C. P. Jayapandian, L. Cui, M. G. Morrical *et al.*, "Scaling up scientific discovery in sleep medicine: the national sleep research resource," *Sleep*, vol. 39, no. 5, pp. 1151–1164, 2016.

[3] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nature reviews. Genetics*, vol. 13, no. 6, p. 395, 2012.

[4] (2017) National sleep research resource (NSRR) launches. [Online]. Available: https://sleep.med.harvard.edu/news/518/NationalSleepResearchResourceNSRRLaunches

[5] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Acm sigmod record*, vol. 22, no. 2. ACM, 1993, pp. 207–216.

[6] J. Hipp, U. Güntzer, and G. Nakhaeizadeh, "Algorithms for association rule mining: a general survey and comparison," *ACM sigkdd explorations newsletter*, vol. 2, no. 1, pp. 58–64, 2000.

[7] S. Kotsiantis and D. Kanellopoulos, "Association rules mining: A recent overview," *GESTS International Transactions on Computer Science and Engineering*, vol. 32, no. 1, pp. 71–82, 2006.

[8] P. Fournier-Viger, C.-W. Wu, and V. S. Tseng, "Mining top-k association rules," in *Canadian Conference on Artificial Intelligence*. Springer, 2012, pp. 61–73.

[9] P. Fournier-Viger and V. S. Tseng, "Mining top-k non-redundant association rules." Springer, 2012.

[10] D. Hristovski, J. Stare, B. Peterlin, and S. Dzeroski, "Supporting discovery in medicine by association rule mining in medline and UMLS," *Studies in health technology and informatics*, no. 2, pp. 1344–1348, 2001.

[11] X. Hu, X. Zhang, I. Yoo, X. Wang, and J. Feng, "Mining hidden connections among biomedical concepts from disjoint biomedical literature sets through semantic-based association rule," *International Journal of Intelligent Systems*, vol. 25, no. 2, pp. 207–223, 2010.

[12] C. Ordonez, E. Omiecinski, L. De Braal, C. A. Santana, N. Ezquerra, J. A. Taboada, D. Cooke, E. Krawczynska, and E. V. Garcia, "Mining constrained association rules to predict heart disease," in *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*. IEEE, 2001, pp. 433–440.

[13] C. Ordonez, "Association rule discovery with the train and test approach for heart disease prediction," *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 2, pp. 334–343, 2006.