Association Rule and Frequent Pattern Mining using the OMOP-CDM

Solomon Ioannou, Egill Fridgeirsson, Jan Kors, Peter Rijnbeek

Background

Data mining tasks aim to extract and analyze information to support decision making [1]. This includes pattern mining, dating back to the early 1990's. Initially, stated as a problem in the 'market basket' domain, pattern mining aims to discover structure and correlations in databases. Applying such methods to observational health data seems promising to reveal interesting and sometimes unexpected patterns [2]. For example, association rule mining aims to answer the question, 'Given a cohort of patients, which concepts are most likely to occur together?' It could be used to measure the association between two or more concepts from any domain in the Common Data Model (CDM), such as conditions, drugs, procedures, etc. Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time.

Parameter	Description
minimum support	threshold for the minimum number of patients that should have the concept set in their medical history, e.g., {obesity, diabetes}
minimum confidence	threshold for determining how often the left side of the rule occurs together with the right side, e.g., {obesity, diabetes} -> {heart failure}

Another example are frequent pattern mining methods that take into account the chronological ordering of concepts. These methods can be used to answer the question, "What are the most frequent sequences of concepts observed in a cohort of patients?" Frequent patterns are required to satisfy minimum support.

Here we present ongoing work on the R package AssociationRuleMining, a framework to perform association rule and frequent pattern mining analysis using data in the OMOP-CDM. The framework provides an opportunity to assess the temporal structures of the medical history of patients which can be used to characterize patients or can be used in patient-level prediction.

Methods

The AssociationRuleMining R package makes use of the open-source SPMF Java library by Phillippe Fournier-Viger [3] that incorporates a large number of association rule and frequent pattern mining algorithms, e.g., "Apriori" [4], "Eclat" [5], and "FP-Growth" [6] for mining highly associated sets of concepts, and "SPADE" [7], "Clasp" [8], and "Prefixspan" [9] for mining frequent patterns.

The AssociationRuleMining Package is fully integrated in the OHDSI framework using DatabaseConnector [10] and FeatureExtraction [11], and runs on any defined cohort. The resultant frequent patterns can be automatically added as custom covariates for use with other OHDSI packages, such as PatientLevelPrediction [12].

Results

After execution, the generated association rules or frequent patterns can be explored in R. Depending on the size of the cohort and the settings, the number of extracted patterns can be very large. We are therefore working on methods to visualize the results interactively. We are exploring interactive networks (Figure 1) to visualize rules and interactive Sankey diagrams to visualize frequent patterns (Figure 2).

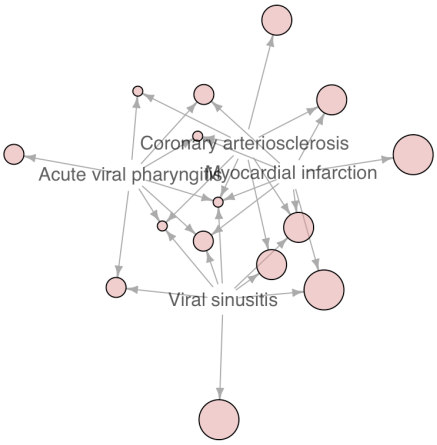

Figure 1: Top 15 rules with highest support in a myocardial infarction cohort.

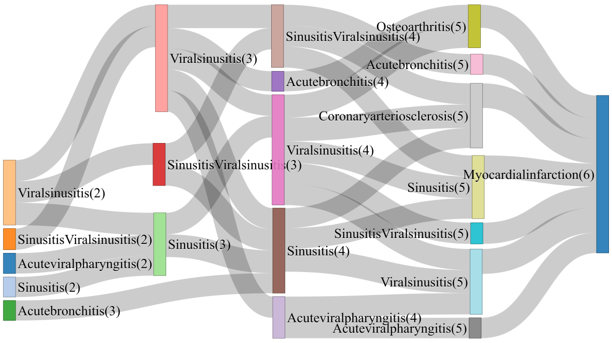

Figure 2: Sankey diagram showing frequent patterns in a myocardial infarction cohort.

Conclusions

Our ultimate aim is to assess the value of different association rule and frequent pattern methods for characterizing patients, and as potential predictors in prediction problems. Therefore, we will further develop this R Package in the upcoming months and will evaluate this in a number of clinical problems.

References

[1] Fournier-Viger P, Lin JCW, Vo B, Chi TT, Zhang J, Le HB. A survey of itemset mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2017;7:1207. https://doi.org/10.1002/widm.1207.

[2] Fournier-Viger P, Chun J, Lin W, Kiran RU, Koh YS, Thomas R. A survey of sequential pattern mining. Data Science and Pattern Recognition 2017;1:54–77.

[3] Fournier-Viger P, Lin JCW, Gomariz A, Gueniche T, Soltani A, Deng Z, et al. The SPMF open-source data mining library version 2. Lecture Notes in Computer Science 2016;9853 LNCS:36–40. https://doi.org/10.1007/978-3-319-46131-1_8.

[4] Agrawal R, Srikant R. Fast algorithms for mining association rules. Proceedings of the 20th International Conference on Very Large Data Bases, VLDM 1994:487–99.

[5] Zaki MJ. Scalable algorithms for association mining. IEEE Transactions on Knowledge and Data Engineering 2000;12:372–90. https://doi.org/10.1109/69.846291.

[6] Han J, Pei J, Yin Y, Mao R. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Mining and Knowledge Discovery 2004;8:53–87. https://doi.org/10.1023/B:DAMI.0000005258.31418.83.

[7] Zaki MJ. SPADE: An efficient algorithm for mining frequent sequences. Machine Learning 2001;42:31–60. https://doi.org/10.1023/A:1007652502315.

[8] Gomariz A, Campos M, Marin R, Goethals B. ClaSP: An efficient algorithm for mining frequent closed sequences. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7818 LNAI, Springer, Berlin, Heidelberg; 2013, p. 50–61. https://doi.org/10.1007/978-3-642-37453-1_5.

[9] Pei J, Han J, Mortazavi-Asl B, Wang J, Pinto H, Chen Q, et al. Mining sequential patterns by pattern-growth: The prefixspan approach. IEEE Transactions on Knowledge and Data Engineering 2004;16:1424–40. https://doi.org/10.1109/TKDE.2004.77.

[10] Schuemie M, Suchard M. DatabaseConnector: Connecting to various database platforms. 2021.

[11] Schuemie M, Suchard M, Ryan P, Reps J. FeatureExtraction: Generating features for a cohort. 2020.

[12] Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. Journal of the American Medical Informatics Association 2018;25:969–75. https://doi.org/10.1093/jamia/ocy032.