Humboldt Universität zu Berlin

Bachelor's Thesis Exposé

Working Title:

Association Rule Mining on medical discharge summaries

Author: Konstantin Böttcher
Reviewer: Prof. Dr. Ulf Leser
Prof. Dr. Niels Pinkwart

1 Background

Data mining procedures can contribute useful information for the medical sector. Knowledge about relations between the vast amounts of symptoms and diseases could help generate diagnostic hypotheses about the condition of a patient or to find mistakes in the medical records. Therefore it seems promising to utilize the precise data already recorded for treatment, as far as this is possible under consideration of privacy rights.

One way to approach this task was introduced by Agrawal et al [1] in the year 1993 and is called Association Rule Mining, short ARM. Initially used to find relations between products (items) sold in retail, ARM finds sets of items, that appear frequently together in transactions and generates rules based on these item sets. These rules make predictions about which items are likely to be part of a transaction, if other items are already.

In the context of medical data, ARM has already been adopted for different purposes. Abdullah et al used a modified version of ARM to find rules for a given item set based on medical billing data [9]. Séverac et al produced a set of rules which linked pathological conditions with medications [10]. Stiloul et al applied ARM on data containing conditions of diabetes patients [11] and Nahar et al examined factors for heart diseases by using ARM [12].

This thesis uses ARM to analyze symptoms and illnesses as reported in discharge summaries, as already done by Doddi et al, where ARM was used on randomly sampled transactions [13] and by Kost et al, who examined diseases co-occurrence on hierarchically structured data [14]. These discharge summaries are created at the end of a hospital admission and consist, alongside a textual description, of a list of codes which translate to observed medical characteristics of the patient. Every discharge summary is interpreted as a transaction and every medical code contained in a discharge summary as an item. The resulting association rules found by ARM will suggest characteristics that appear often in combination or as consequence of other ones.

2 Goal

The goal of this bachelor's thesis is to first use a specialized form of ARM, called Generalized Association Rule Mining, to find relations between medical codes in discharge summaries and second to compare these findings with the results of a thesis by Jonathan Bräuer [5], which attended to predict medical codes by processing diagnoses written in natural language. Key interest is the amount and kind of overlap between false-positive codes, generated from natural language and corresponding findings using association rule mining on the same discharge summaries, thus examining differences between the actual codes within discharge summaries in respect to the generated codes.

3 Dataset and Structure

The dataset used in this thesis is the MIMIC III database [6], which consists of medical data gathered at the Beth Israel Deaconess Medical Center in the timespan from 2001 to 2012. It contains data for over 58.000 hospital admissions. Every admission is connected to a discharge summary, containing medical codes using the ICD9-CM classification system. In total there are around 650.000 codes assigned to the discharge summaries, thus every discharge summary contains on average 11 codes.

ICD9-CM is structured as a taxonomy with four subtrees [8], one containing symptoms and diseases, a second one containing medical procedures, a third containing not directly health related characteristics, as pregnancy and the last one mainly containing injuries. These four subtrees are divided into more specific subcategories and on the lowest levels in sub specifications of diseases, injures and procedures. For this thesis only the subtree containing diseases and symptoms is relevant. The lower level codes in this subtree consists of three digit numbers describing the diseases and up to two more digits for sub specifications. For example 001 meaning “Cholera” and 001.0 meaning “Cholera due to vibrio cholera”. The higher levels of the taxonomy are abstract categories, represented by two three digit numbers showing the range of codes that are in that category, e.g. 001-009 for “Intestinal Infectious Diseases”. These high level codes are not used on discharge summaries.

Entries in the MIMIC database can originate from different, though not all levels of the ICD9-CM taxonomy and are coded as specific as the observed medical characteristic could be determined.

4 Association Rule Mining

The input for ARM is a set of transactions. Every transactions is a set of items. The goal of ARM is to find meaningful relations between items, in form of so called association rules. These rules are structured as tuples of two item sets, the antecedent or left hand side and the consequent or right hand side of the rule. To find meaningful rules Agrawal et al proposed two measures [1]. First the support of a rule, given by the amount of occurrences of the combined item set of the rule in the transaction database and second the confidence, determined by the occurrences of the consequent divided by the support. Rules that don’t meet a minimum threshold for both these criteria are pruned from the results.

The original approach for ARM only produces rules for items with no hierarchical structure. However the given dataset is based on the ICD9-CM standard, which incorporates a taxonomy with growing specificity on every level. Therefore the original approach of ARM could not derive rules for generalizations, also called ancestors, of items. If for example the transaction database would only contain entries for 001.1 ARM would not be able to generate rules with the code 001. Furthermore if a transaction contains a descendants of an item this entry would not be counted to the support of the ancestor. Imagine a database that contains some transactions with code 001 and some with code 001.0. Further suppose that both codes appear less frequent than the minimum support, but combined appear more frequent. ARM would not find any rules for 001, which is obviously not correct as every entry for 001.0 also should be counted as an entry for 001.

Therefore a Generalized Association Rule Mining approach will be applied for this thesis. First introduced by Srikant et al in 1995 [4], Generalized Association Rule Mining, or GARM, allows the usage of hierarchical structured data. It incorporates the taxonomy the items are arranged in and is able to find rules for items originating from different levels of the hierarchy. This is achieved by expanding all transactions with the ancestors for every item contained in the transaction. For example a transaction with the entry 001.0 (Cholera due to vibrio cholera) would be expanded with 001 (cholera). Because the higher levels of the taxonomy are only abstract categories, those will not be considered. Afterwards redundant item sets, containing items and their ancestors will be pruned.

GARM will be implemented using the FP-tax algorithm by Pramudiono et al [3], which builds on the FP-growth algorithm by Han et al [2]. FP-Growth is more efficient than the originally by Agrawal et al introduced Apriori algorithm [1], because it doesn't uses candidate generation, but instead builds a frequent pattern tree to find all wanted item sets. FP-tax expands FP-growth to incorporate the taxonomy, the dataset is based on. For FP-growth the implementation in the SPMF Data-Mining library [7] will be modified to fit FP-tax.

When generating the rules two additional pruning methods, alongside minimum thresholds for support and confidence will be used. First only rules with one element in the consequent will be generated. Because the main interest of the thesis is the comparison with the predicted codes from the Thesis by Bräuer [5], it is not necessary to mine rules with larger consequents. This will be explained in the evaluation section. Second this thesis uses a pruning strategy proposed by Srikant et al in the original paper about GARM [4], to prune non interesting rules. For all rules that have ancestors, meaning a rule containing at least one ancestor of an item from the original rule, it will be checked if this rule has lower support than expected by examining the support of the ancestor-rule. If this is the case the rule can be pruned, because it doesn't contain any new information.

5 Comparison

The baseline for the comparison are the predicted codes from the thesis by Bräuer [5]. It will be carried out by using all codes that were detected as false positive on the corresponding discharge summary. For every code, all rules with this code as the consequent will be selected. Then it will be checked if the antecedent of any of these rules is a subset of the corresponding expanded transaction set, which also contains its items ancestors. If this holds true for at least one rule there is indication that this item should have been part of the discharge summary and was forgotten.

In a second step it will be examined if there are rules for ancestors or descendants of generated codes, which didn't had a matching rule already. For codes whose descendent has a matching rule it can be argued that the code isn't specific enough and vice versa if there is a rule for an ancestor that the generated code is to specific.

References

[1] Agrawal R., Imielinski T., and Swami A. (1993). Mining association rules between sets of items in large databases. In Proc. of the ACM SIGMOD Conference on Management of Data, pages 207-216.

[2] Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Mining and Knowledge Discovery, 8, pages 53-87. DOI: 10.1023/B:DAMI.0000005258.31418.83

[3] Pramudiono I. and Kitsuregaw M. (2004). FP-tax: tree structure based generalized association rule mining. DMKD '04 Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, pages 60-63.

[4] Srikant R., Agrawal R. (1997). Mining Generalized Association Rules. In Proceeding VLDB '95 Proceedings of the 21th International Conference on Very Large Data Bases, pages 407-419.

[5] Bräuer J. Clinical Entity Recognition for ICD-9 Code Prediction in Clinical Discharge Summaries. Thesis, Humboldt Universität zu Berlin.

[6] MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available from: http://www.nature.com/articles/sdata201635

[7] Fournier-Viger P., Lin C.W., Gomariz A., Gueniche T., Soltani A., Deng Z., Lam H. T. (2016). The SPMF Open-Source Data Mining Library Version 2. Proc. 19th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2016) Part III, Springer LNCS 9853, Â pp. 36-40.

[8] https://bioportal.bioontology.org/ontologies/ICD9CM

[9] Abdullah U., Ahmad J. and Ahmed A. (2008). Analysis of effectiveness of apriori algorithm in medical billing data mining. 4th International Conference on Emerging Technologies, Rawalpindi, 2008, pages 327-331.

[10] Séverac F., Sauleau E.A., Meyer N., Lefèvre H., Nisand G., Jay N. (2015). Non-redundant association rules between diseases and medications: an automated method for knowledge base construction. BMC Medical Informatics and Decision Making (2015), pages 15-29.

[11] Stilou1 S., Bamidis P.D., Maglaveras N., Pappas C. (2001). Mining Association Rules from Clinical Databases: An Intelligent Diagnostic Process in Healthcare. Studies in Health Technology and Informatics 2001, Volume 84:1399-1403

[12] Nahar J., Imama T., Tickle K.S., Chen Y.P. (2013). Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Applications, Volume 40, Issue 4, 2013, pages 1086-1093.

[13] Srinivas Doddi, Achla Marathe, S. S. Ravi, David C. Torney (2001) Discovery of association rules in medical data, Medical Informatics and the Internet in Medicine, 26:1, pages 25-33.

[14] Kost R., Littenberg B., Chen E.S. (2012). Exploring Generalized Association Rule Mining for Disease Co-Occurrences. AMIA Annual Symposium Proceedings. 2012: pages 1284-1293.