SURVEY OF ONTOLOGY ASSOCIATION RULE EXTRACTION FROM GENE ONTOLOGY ANNOTATIONS

R. Jayasri\textsuperscript{a} and S. Saravanan\textsuperscript{b}

\textsuperscript{a}M.Tech. , Department of Computer Science and Engineering, SRM University, Kattankulathur, India.
\textsuperscript{b}Assistant Professor (Sr.G), Department of Computer Science and Engineering, SRM University, Kattankulathur, India.
Email: jayasri13194@gmail.com

Received on 10-11-2016
Accepted on: 29-11-2016

Abstract

Objectives: Recent years, we have used various data mining approaches to get the bio-related information. The main motive of mining algorithm is to mine rules by linking Gene Ontology Terms. Gene Ontology is a constructed depository of notions (GO Terms) that are associated to one or more gene products by annotation process. The concepts behind algorithms are conveyed initially, proceeded by an overview to some of the previous research works.

Methods: Here we discuss various mining algorithms such as Classical Association Rule Mining Algorithm (ARL), Direct Hashing and Pruning Algorithm (DHP), Multi-Ontology data mining at All Levels (MOAL) and Gene Ontology Weighted Association Rule (GO-WAR) that provide us consistent rules. These techniques in addition help to also identify the inner relationships between terms. In specific, ARL and DHP mines ontology association rules, MOAL mines cross ontology association rules that involves different formulation of Support and Confidence for same ontologies and different ontologies and GO-WAR mines cross ontology weighted association rules using intrinsic methods (topology-based) and extrinsic methods (annotation-based) of Information Content.

Results: The result of mining algorithms produce association rules and those rules help in providing assistance to biological researchers and ontology curators.

Conclusion: The performance of mining algorithm is based on the three main factors: completeness, consistency and non-redundancy of the information. The paper aims at giving survey theoretically on some of the existing algorithms.

Keywords: Gene Ontology, Weighted Association Rules, Information Content
Introduction

**Gene ontology** project aims to maintain gene and gene product attributes. The significant goal of GO project is to provide tools for easy access to all aspects of the data provided by the project, and to enable functional interpretation of experimental data using the GO. It develops a set of controlled, constructed ontologies to describe key domains of molecular biology. Gene product attributes and biological sequences are included in it. GO project works on GO terms in the annotation of sequences, genes or gene products in biological databases. It also provides a open standard access of public resource to the ontologies, annotation data sets and software tools. It is developed for use with GO data.

**Ontology**

The Gene Ontology is structured as Directed Acyclic Graph (DAG) [3]. The ontology covers three domains:

**Cellular Component (CC):** It describes locations, at the levels of sub cellular structures and macromolecular complexes. Example: `nuclear inner membrane`, with the meaning `inner envelope`.

**Molecular Function (MF):** It is the genetic activities at the molecular level, such as binding or catalysis.

**Biological Process (BP):** Group of molecular events with a defined beginning and end. It describes biological goals accomplished by one or more ordered assemblies of molecular functions. Example: High-level processes such as `cell death` can have both subtypes, such as `apoptosis`, and sub-processes, such as `apoptotic chromosome condensation`.

**Annotations:** Genome annotation is the information about a gene product and GO annotation terms from the GO ontology. Annotations should be attributed a source which is of computational design or database. In order to support the association between gene product and GO Terms, the annotation must show the type of evidence used. Hence High-quality GO annotations are supported by experimental evidence which is provided by the source cite. A standard set of evidence codes always qualify the annotations based on experimental determinations. Evidence Code comes from Evidence Code Ontology. The analysis of annotated data is therefore an important opportunity for bioinformatics since it may represent a source of biologically meaningful knowledge.

**Materials and Methods**

**Classical Association Rule Learning Algorithm (ARL)**

This technique is developed by Faria et al [19]. It discovers redundant or inconsistent annotations. The main focus of this algorithm is to find the implicit relationships between terms. If annotation is redundant, true path rule is applied. This will avoid the redundant relationship. After that, the relationships that share descendants are removed.
Direct Hashing and Pruning Algorithm (DHP)

Large itemset generation and TDB size reduction is the main motive of DHP Algorithm [6]. The magnitude of candidate itemsets of DHP is comparably less when compared to other existing algorithms. In case of TDB, best pruning techniques are implemented to reduce it’s size.

Multi-Ontology data mining at All Levels Algorithm (MOAL)

MOAL Algorithm mines cross ontology association rules [12]. This algorithm uses modified apriori algorithm as their mining strategy and the items are non-weighted. A modified version of Apriori algorithm is used for mining rules [16] where the items are non-weighted. It uses either MF Ontology or multi Ontology. As the Ontology implies, multi-ontology rules are categorized as Cross-Ontology Multi-Level rules (CO_ML) where ontologies are different in two terms of a rule and Same-Ontology Multi Level rules (SO_ML) where Ontologies are same in two terms of a rule. They are categorized since the formulation of MO_Support and MO_Confidence are different for CO_ML and SO_ML. Another approach of Manda et al [22] is used in ARL from Gene Ontology.

Gene Ontology-Weighted Association Rule Algorithm (GO-WAR)

GO-WAR is the algorithm that mines weighted association rules [2] from GO. It uses an optimized FP-Growth algorithm. The goal of this algorithm is same as that of the remaining association rule mining algorithm (finding out relationship among the terms). GO-WAR is a general approach that can be applied to any type of data where the goal is to discover relationships among the terms. The steps of GO-WAR Algorithm include transformation, mining and pruning. Transformation converts input data file in a Transaction Database, Mining mines rules using Weighted Support value by building FP-Tree and Pruning removes non informatary rules using Confidence and Information Content values. Information Content notion are categorized as intrinsic and extrinsic method. Intrinsic approach considers only the structural information (extracted from ontology). Intrinsic IC is calculated using Sanchez method [21].

Results and Survey:

ARL

Basically, general terms give changes in the results during the extraction of Association Rules (AR). The main drawback of this approach is, Faria et al. didn’t deal with the general terms and different sources of production of GO annotations. Inorder to filter those terms, they proposed specificity formulation called Information Content (IC) [23] but they produce candidate rules with low IC when used on annotated data.
DHP

To access the performance of DHP, the Apriori Algorithm and the DHP Algorithm have been compared. The execution time of DHP is more. It is same as that of Apriori Algorithm in the generation of large itemsets. The variation is that DHP scans the smaller transaction database and uses hash table for the generation of second candidate table.

MOAL

MOAL uses modified version of Apriori algorithm [16] using non-weighted items to mine rules. The algorithm learns association rules using the standard minimum Support and Confidence measures, adding to each rule a p-value threshold computed using the Chi-square test. The drawback is, chi-square test does not give us much information about the strength of the relationship. The another major disadvantage is the time complexity of the algorithm increases. This is because of the frequent access of database.

GO-WAR

The first factor is the size of the input dataset where it requires more memory when analyzing the dataset. The FP-Tree dimension varies on the basis of weighted values of Support and Confidence. High values reduce the FP-Tree dimension and low value increase the FP-Tree dimension.

Table 1: Overall Survey on Existing Algorithms

<table>
<thead>
<tr>
<th>S</th>
<th>Author Name</th>
<th>Title</th>
<th>Algorithm/Formulation</th>
<th>Drawbacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>S. P. Jong, S.C Ming and S. Y. Philip</td>
<td>An effective hash based algorithm for mining association rules.</td>
<td>Direct Hashing and Pruning Algorithm (DHP)</td>
<td>Execution time is more for small transactions</td>
</tr>
<tr>
<td>2</td>
<td>Han, Pei, Y. Yin and R. Mao</td>
<td>Mining Frequent Patterns without Candidate Generation:A Frequent-Pattern Tree Approach</td>
<td>FP-Tree Approach</td>
<td>Important rules are missed while mining</td>
</tr>
<tr>
<td>3</td>
<td>P. Manda, S. Ozkan, H. Wang, F. McCarthy and</td>
<td>Cross-ontology multi-level association rule mining in the</td>
<td>Multi-Ontology data mining at All Levels Algorithm (MOAL)</td>
<td>Important rules such as binary rules and extended rules are missed while</td>
</tr>
</tbody>
</table>

<p>|</p>
<table>
<thead>
<tr>
<th>Page</th>
<th>S. M. Bridges</th>
<th>gene ontology</th>
<th>mining</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>G. Agapito, M. Milano, P. H. Guzzi, and M. Cannataro</td>
<td>Improving annotation quality in gene ontology by mining cross ontology weighted association rules</td>
<td>Basic GO-WAR</td>
<td>Rules generated are less</td>
</tr>
<tr>
<td>5</td>
<td>D. Faria, A. Schlicker, C. Pesquita, H. Bastos, A. E. N. Ferreira, M. Albrecht, and A. O. Falco</td>
<td>Mining go annotations for improving annotation consistency</td>
<td>Association Rule Learning Algorithm</td>
<td>1. It can handle only MF Ontology. 2. It uses traditional apriori algorithm for mining rules. 3. It produces candidate rules with low IC. 4. They do not consider Weighted Support.</td>
</tr>
<tr>
<td>7</td>
<td>Li Qingzhong, Wang Haiyang and Yan Zhongmin</td>
<td>Efficient mining of association rules by reducing the number of passes over the</td>
<td>Mining Algorithm</td>
<td>Time Complexity is more</td>
</tr>
</tbody>
</table>
Table

<table>
<thead>
<tr>
<th>Page</th>
<th>Authors</th>
<th>Title</th>
<th>Algorithm</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>R. Agarwal and R. Srikant</td>
<td>Fast algorithms for Mining Association Rules in Large Databases</td>
<td>Fast Algorithms</td>
<td>1. Database is not effectively trimmed 2. Occurrence of item sets are more</td>
</tr>
<tr>
<td>9</td>
<td>P. Resnik</td>
<td>Using information content to evaluate semantic similarity in a taxonomy</td>
<td>Resnik IC Formulation (Intrinsic and Extrinsic Approach)</td>
<td>It takes value in the extended intervals</td>
</tr>
</tbody>
</table>

Conclusion

The above novel algorithms are used for mining weighted association rules. ARL algorithm is the weighted mining technique that gives rules by showing relations between the GO terms using classical Support and classical Confidence. It also surveys that the incomplete informations, the redundant informations and the inconsistent informations affect the performance of association rule mining algorithms. The further advancement is the MOAL algorithm that mines cross ontology association rule. This approach is processed based on MO Support and MO Confidence where its formulation varies for SO_ML and CO_ML. DHP Algorithm is also called as trimming technique that lessens the transaction count. GO_WAR is the weighted association rule mining algorithm that mines rules between either same ontologies or different ontologies. Hence this approach seems to be more adaptable. It avoids manual intercession.

References


7. Han, Pei, Y Yin and R Mao, “Mining Frequent Patterns without Candidate Generation: A Frequent- Pattern Tree Approach. Data Mining and Knowledge Discovery”, 2004.


