UNIVERSITI TEKNOLOGI MARA

RANKING-BASED PRUNING AND WEIGHTED SUPPORT MODEL FOR GENE ASSOCIATION IN FREQUENT ITEMSETS

SOFIANITA MUTALIB

Thesis submitted in fulfilment
of the requirements for the degree of
Doctor of Philosophy
(Information Technology and Quantitative Sciences)

Faculty of Computer and Mathematical Sciences

August 2019

AUTHOR'S DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Teknologi MARA. It is original and is the results of my own work, unless otherwise indicated or acknowledged as referenced work, this thesis has not been submitted to any other academic institution or non-academic institution for any degree or qualification.

I, hereby, acknowledge that I have been supplied with the Academic Rules and Regulations for Post Graduate, Universiti Teknologi MARA, regulating the conduct of my study and research.

Name of Student:	:	Sofianita Mutalib
Student I.D. No.	:	2010321805
Programme	:	Doctor of Philosophy (Information Technology and Quantitative Sciences)
Faculty	:	Computer and Mathematical Sciences
Thesis Title	:	Ranking-Based Pruning and Weighted Support Model for Gene Association in Frequent Itemsets

Signature of Student	:	....
Date	:	August 2019

ABSTRACT

Biological domain is one of the critical areas that always seek for useful knowledge and patterns observed through available methods, including data mining. One of genomic benchmark data sources is from Genome Wide Association Studies (GWAS), which uses a set of genetic variants, namely Single Nucleotide Polymorphisms (SNPs), in different individuals to observe the association of the variants with a particular trait. Usually, the association test in GWAS is done by finding the risk measure of each of the SNPs separately. But many of the variants and its effect remain a mystery which has high potential of knowledge discovery especially for complex diseases. The aim of the research is to develop an improved method for processing information and to find the relationship between genetic variants and disease with in-depth interpretation. Therefore, this research attempts to investigate the association between genetic variants to diseases, and thus propose a method that can identify multiple SNPs combination to form an association using Frequent Itemset Mining (FIM). Five main stages of methodology in this research are, data understanding, data representation and pruning items, FIM and analysis and validation of knowledge. This thesis elaborates a set of crucial tasks in FIM for GWAS datasets. It proposes a strategy of Ranking-based Pruning of Items (RPI) for SNPs. Next, the development of Weighted Support Model (WSM) was done to search for interesting itemsets. The measurement used are Information Gain for ranking to prune items and Weighted Support for interestingness of itemset. High dimensional dataset presented by SNPs confirmed the reason to apply row enumeration strategy algorithm to mine frequent closed itemsets. It is found that SNPs with known risks to Type 2 Diabetes Mellitus (T2DM) occur in low support values, that cause the process of searching frequent itemsets to be repeated many times until the low support values are retrieved. The implementation of WSM with Odds Ratio (OR) values, gives visibility of these itemsets as higher *weighted support* value. Finally, the validation for interestingness of produced itemsets is through the integration of available and relevant biological information with scrutinization of an expert as presented in the Descriptive Gene set Analysis (DGA). The information found in the itemsets concluded that the identified SNPs interact with other variants in the chain of T2DM. The scope of the work is using two most commonly chromosomes of T2DM studied, which are Chromosome 11 and 16. The results show that the itemsets with the T2DM risk variants were found within the *support* values of 40 to 48, and after the RPI and WSM are applied, the *weighted support* value increases to 50 and 97 within significant number of SNPs. These results show that RPI-WSM is able to solve the huge dataset problem and low support value problem in FIM. In addition, to improve the interpretation, each itemset is presented as combination of genes in DGA with gene annotation information, that supplies scientist with further valuing patterns. RPI, WSM and DGA are the contributions of the research and significant in discovering potential new knowledge and complimenting research by scientists to perform further validations. The study could also contribute to the advancement in healthcare and digital genome market, which focuses on developing healthier society through monitoring and early protection of any threats, especially of chronic diseases such as T2DM through personalized treatment or medicine.

TABLE OF CONTENTS