Weighted Frequent Itemset of SNPs in Genome Wide Studies

Sofianita Mutalib, Azlinah Mohamed, Shuzlina Abdul-Rahman, and Norlaila Mustafa

***Abstract*—Genome wide association study (GWAS) is a study to investigate the correlations between genetic variants and traits. GWAS normally focus on the associations between single-nucleotide polymorphisms (SNPs) and traits like major human diseases. Generally, GWAS uses standard statistical tests on each SNP to capture main the genetic effects. However, the association is done between a single SNP and the trait. This study make use the whole sets of available SNPs in GWAS, data mining approach is applied to associate more than one SNPs to traits. In general, this will complement the GWAS to help understand complex diseases. This paper presents a proposed frequent itemset mining with weights to discover important sets of SNPs that are associated with diabetes. The purpose of using weights is to mine SNPs that might be less frequent but important in the study of diabetes. The approach consists of three stages: first, reduction of feature space and testing them through classifiers; second, the selection of informative SNPs through allelic testing; then, weight assignment for the selected SNPs; and third, itemset mining and gene analysis. The proposed approach has proven to be effective by helping to discover genes that have associated with the risk of diabetes. These patterns could be used as a set of significant information extracted by mining genetic variants in any particular SNP.**

***Index Terms*—Diabetes, feature selection, frequent itemset mining, single nucleotide polymorphism, weight.**

I. Introduction

The biological data are abundant and information on the diversity of genome sizes range anywhere from megabytes to terabytes. Recently, a Single Nucleotide Polymorphism (SNP) as a unit of genetic variations in deoxyribonucleic acid (DNA) has caught much attention as it is associated with complex diseases. These data are produced in different formats and in different files, and genome wide data studies collect genotype and phenotype data including SNPs. There are various methods to analyze the data, the choice of which depends on the problem and type of knowledge that is expected to be meaningful to the scientists.

The growing amount of data in the healthcare industry promotes the hidden knowledge discovery that could be mined in the massive data. The industry also is aware the knowledge can benefit them is so many ways and the deep analysis can be done within their databases with the availability of various data mining tools and techniques. The analysis has to be done in a correct way with the involvement of medical knowledge and close cooperation between data analysis experts and physicians. In later stages, the discovery of intended knowledge can be integrated in existing systems for many purposes such as in research, diagnosis, and treatment of critical diseases. It can be used both by physicians and as a part of systems, such as expert and knowledge management systems. Raw biological and medical data by nature are heterogeneous, so these in itself bring more challenges in data mining. The data structure can contain various data of different data types, like integer, float, string and nominal. Simultaneously, the amount of data is growing and contributing to the volume of the data.

There are a lot of studies that have been carried out to investigate the role of an individual's genetic factor in determining an exposure to disease, diagnosis or personalized medicine [1]-[4]. Genomics and proteomics research have found new associations between genetic variants, and these have high potentials in improving healthcare practices. One of the most challenging tasks of current bioinformatics studies, including medical and genomic research, is to associate genotype (genetic) and phenotype (traits) information. The genotype information is normally generated by high-throughput genotyping technologies and the data can be retrieved from private databases or publicly accessed from available repositories. However, due to its data format and enormous feature space SNP analysis make it a complicated task and machine learning techniques could be appropriate to be applied in achieving the stipulated goal.

Currently, many common diseases are rigorously under investigation for its genetic factors, implications of the diseases, as well as the treatment. For most complex diseases, the underlying genetic factors remain largely unknown. Complex diseases such as cancer, diabetes and heart disease result from a complex interaction of genetic and environmental factors, which means the diseases are caused by multitude factors. Nevertheless, the pace of progress in the development of genomic research has achieved a great number of successes in discovering susceptible genes and genetic variants. The impact to genetic profiling personalized medicine is high. Genetic profiling is the result of simultaneous testing at multiple genetic loci [5], [6].

We, therefore in this particular paper, explore methods to find the important genetic variants, which is referred to Single Nucleotide Polymorphisms (SNPs). In the singular pursuit of the study, we applied feature selection methods with accuracy evaluation and also, ranked the selected SNPs with regards its score using statistical measure - odd ratio, the measure of an

Manuscript received May 5, 2018; revised July 1, 2018.

Sofianita Mutalib, Azlinah Mohamed, and Shuzlina Abdul-Rahman are with the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor Malaysia (e-mail: sofi@tmsk.uitm.edu.my, azlinah@tmsk.uitm.edu.my and shuzlina@tmsk.uitm.edu.my).

Norlaila Mustafa is with the Medical Department, Faculty of Medicine, Hospital Canselor Tuanku Muhriz, Jalan Yaacob Latif, Bandar Tun Razak, Universiti Kebangsaan Malaysia, 56000 Cheras, Kuala Lumpur, Malaysia (e-mail: norlaila@ppukm.ukm.edu.my).

doi: 10.18178/ijmlc.2018.8.4.704

exposure and an outcome. Finally, the frequent itemsets are discovered through mining processes and the implications of weight are presented. The rest of the paper is organized as follows. The summarization of related studies in genetic variants is given in Section II. Section III describes methods and experimental setup. The results and discussions are discussed in Section IV. Finally, the conclusions are presented in Section V.

II. Related Studies

Genetic variants have made association studies a powerful approach for mapping complex-disease genes by conducting studies based on the whole genome [7], [8]. Although conducting association studies based on genetic variants is practical, it is not practical or statistically feasible to genotype and test all SNPs in the genome in conducting association studies. The GWAS data can be considered a high dimensional dataset, in which the data is very dense and scarce, with a large portion of irrelevant SNPs to the disease. So, to overcome irrelevant SNPs, selection and identification of SNPs has become among the most important tasks in GWAS [9], [10]. Some studies have also listed feature reduction methods that are applicable to genome wide studies. The feature selection/reduction method could be implemented together with classifiers in order to experiment several sets of features that produce good results. For example, Moore *et al.* [6] included filter or wrapper algorithms in GWAS analyses framework. Table I below shows several studies of feature selection for classification of SNPs data. With the success of the studies, it motivates us to further investigate classification methods in GWAS.

TABLE I: Feature Selection Approach in GWAS

Authors	SNP Selector	Methods
Calabria *et al.* [11]	SNPRanker	Core scoring function
Dai *et al*. [12]	SHARE	Adaptive Regression
He and Lin [13]	GWASelect	Regression models
Waddell *et al.*, 2005 [9]	Statistical measure Select the top 10% (300) for Multiple Myeloma using info gain and testing using leave-one-out cross-validation an accuracy estimate of 71%,	Support Vector Machine, Decision Tree and Naives Bayes
Liu *et al*., 2013 [14]	Filter method, logistic regression coupled with likelihood ratio test, with 528 173 SNPs, using 50 SNPs and 200 SNPs in Breast Cancer	Support Vector Machine

The statistical measures applied by Waddell [9] are commonly used in feature selection. Information gain (IG) tells us how important a given attribute/item in the dataset is. Information gain is the expected reduction in entropy caused by partitioning the examples according to a given attribute. The entropy of a data set is:

$$\text{Information\_Gain(feature)} = -p\log p-(1-p)\log(1-p) \quad (1)$$

Meanwhile, Gain Ratio (GR) method is an enhancement of the IG method which can overcome the bias in multi-value attributes. In contrast to IG, which measures by selecting attributes with multi-value, the GR method is intended to maximize the feature's information gain and minimize the number of its value simultaneously. The GR formula is given as:

$$\text{Gain\_Ratio(feature)} = \frac{\text{Information\_Gain(feature)}}{\text{SplitInfo}(S)} \quad (2)$$

$$SplitInfo(S) = -\sum \frac{|S_i|}{|S|}\log_2 \frac{|S_i|}{|S|} \quad (3)$$

A number of methods for analyzing the susceptibility of SNPs in GWAS have been proposed in the literature, where each SNP is analyzed individually [15]. The conventional single SNP testing did not provide a good solution for disease prediction and multiple SNP analyses could provide a better solution. In performing the multiple SNP analyses, machine learning or data mining method could be applied and the solution would discover more combinations of SNPs in identifying risk factors. However, it is found that only a small portion of the SNPs have main effects on the complex disease traits, but most of the SNPs have shown little penetrance individually. On the other hand, many common diseases in humans have been shown to be caused by complex interactions among multiple SNPs. This is known as multilocus interactions [16], [17]. Due to the phenomenon, machine learning and data mining methods have been proposed to be applied in associating multiple SNPs, including classification, clustering and association analysis [18]-[20]. However, this paper limits the review to frequent itemset mining that has its origin from rule mining method, as illustrated in Fig. 1 below.

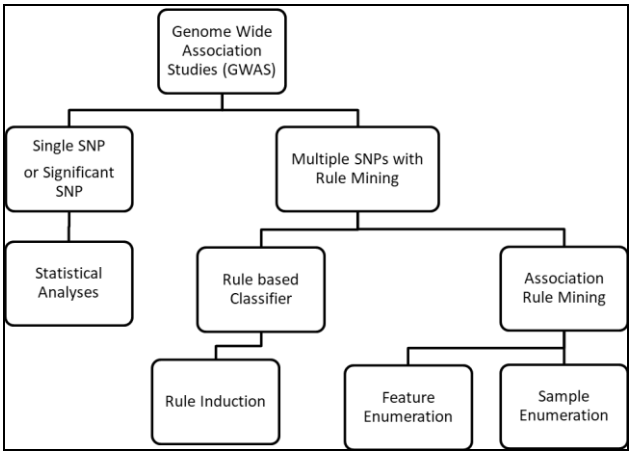

Fig. 1. GWAS and SNP association tests.

Frequent itemset mining (FIM) has been extensively studied in data mining fields. Apriori algorithm was originally designed for supermarket data mining, and the itemsets or rules found can be used for marketing and product selling purposes. It is a classical algorithm for discovering itemsets and generating rules by Agrawal [20]. Given customers' transactions, the algorithm can identify sets of items that frequently co-exist in transactions. For example, by knowing that customers usually purchase any two items together, a shop owner can put the items next to each other in the store. The main processes in Apriori as association rule discovery in data mining are:

1) Finding all frequent itemsets with certain support value

or threshold.

2) Generating strong association rules from the frequent itemsets that meet confidence threshold.

Several studies [21]-[23] have explored the association of genetic variants to diseases using Apriori-based algorithm.

Other than that, many studies have been carried out to propose algorithms that search for frequent closed itemsets with various data structures and operations, in order to improve efficiency [24]-[26]. Based on these literatures, Apriori-based algorithms are most likely inefficient for large numbers of items due to the candidate generation process, apart from consuming memory space. There are two ways of enumerating the frequent itemsets, namely feature or column enumeration, and sample or row enumeration. CHARM and FP-growth algorithms [27] are feature/column enumeration approach and the algorithms are suitable for datasets with small number of features. The alternative approach, which is row enumeration algorithm, intersects transactions to find closed frequent itemsets. The example of row enumeration algorithms are Carpenter [25] and TD-Close [26]. In order to discover the itemset, searching strategy is to be applied in the row enumeration algorithm, normally bottom-up search strategy or top-down search would be the options.

Since Apriori is being proposed, more studies have elaborated and enhanced this algorithm. However, two main bottlenecks exist, which are a huge set of patterns are generated and most of them are redundant. To overcome these problems, three main approaches have been developed. First, efficient algorithms by tree-based and without candidate generation [27] have been developed by using improved search strategies, data formats and structures, and also traversal techniques. Second, closed FIM algorithms proposed [28]-[30] to reduce redundant patterns and extract smaller number of frequent itemsets, The subsets of itemsets, called closed frequent itemset is fewer than frequent itemset. Third, constraint based pattern mining algorithms [31]-[33] have been suggested to address how to reduce uninteresting patterns in terms of users' focus. In the third approach, the mining process focuses on the users' interest by pushing the constraint. As one of constraints, itemset mining approaches with a weight constraints [34]-[37] have been suggested with most of the weighted association rule mining algorithms adopting an Apriori algorithm based on the downward closure property.

A weight of each item is given to reflect the importance of each item in the transaction database in weighted FIM. A weighted pattern is used to represent a set of weighted items and a weight of the itemset is an average value of weights of items within the itemset. A weighted support of a itemset is utilized to prune weighted infrequent patterns and it is the resultant value of multiplying the itemset's support with the weight of the itemset. A weighted transaction database, *D* is defined as follows: *D* comprises of a set of transactions, a set of item, *I* and a set of positive weights corresponding to each item in *I*. For example, consider a dataset comprising of five transactions $T = \{t_1,\ldots, t_5\}$, and six items $I = \{ i_1, i_2, \ldots , i_6\}$. The weights of these items are given as, $W = \{0.1, 0.5, 0.4, 0.2, 0.8\}$, which 0.1 is for i_1, 0.5 is for i_2, 0.4 is for i_3, 0.2 is for i_4, and 0.8 is for i_5 tively. The discovery of weighted itemsets, is executed by finding the occurrence of an itemset in a given transaction and it is weighted by the weight of its least interesting item. Some of advantages gained from weighted itemset mining are:

1) The algorithm can be used to discover the less frequent and weighted itemset.
2) The algorithm can be used for mining high utility itemsets or important itemsets.

Next, the section describes the method to adapt weighted itemset mining in SNPs dataset, which is sparse and dense.

III. Method and Experiment Setup

In this paper, we present a method to associate genes in diabetes that integrates weighted and FIM with feature selection. Given the proposed method, we need to perform SNP data representation and SNP selection as part of preparing the dataset for FIM with importance. The tasks in this particular research includes selecting a set of interesting SNPs from a pool of possible SNPs, evaluate the subset of SNPs in disease prediction using classification model, weighting the importance of each feature (SNP) on a given disease model and finally, evaluate the itemsets found based on the support and weighted support value.

The problem of discovering diabetes association from a set of genetic variants is transformed into finding frequent patterns of association of genetic variants. The data source used in this experiment is real data sets from WTCCC generated using Affymetrix GeneChop 500K Mapping Array Set. The datasets were downloaded from The European Bioinformatics Institute website (EMBL-EBI) [38] and the experiment in this paper was done on Chromosome 16 only. The selected disease, which is T2D, has a number of increasing cases which have been reported in many countries, [38]-[41] and the side effects of the disease is the possible complication to internal organs and also pre-mature death.

A. Data Preparation

TABLE II: Data Representation for Five SNPs and Six Samples

Individual	i_1,	i_2,	i_3,	i_4,	i_5
x_1	$i_{1=AC}$,	$i_{2=GT}$,	$i_{3=CT}$,	$i_{4=AA}$,	$i_{5=CC}$
x_2	$i_{1=AC}$,	$i_{2=GT}$,	$i_{3=TT}$,	$i_{4=AG}$,	$i_{5=CT}$
x_3	$i_{1=AC}$,	$i_{2=GT}$,	$i_{3=TT}$,	$i_{4=AG}$,	$i_{5=CC}$
x_4	$i_{1=AC}$,	$i_{2=GG}$,	$i_{3=CT}$,	$i_{4=AA}$,	$i_{5=CT}$
x_5	$i_{1=AA}$,	$i_{2=GT}$,	$i_{3=CC}$,	$i_{4=AG}$,	$i_{5=CC}$
x_6	$i_{1=AA}$,	$i_{2=GG}$,	$i_{3=CT}$,	$i_{4=AA}$,	$i_{5=CT}$

Tped file data were preprocessed and the data is tabulated into an appropriate format. SNP data were replaced with dbSNP information and transposed several times in different steps. The SNPs datasets are treated as categorical data. Let $X = \{x_1, x_2, \ldots, x_m\}$ be a set of samples and $I = \{i_1, i_2, \ldots, i_k\}$ be a set of k elements called items of sample. The set of individuals, *X*, from GWAS is stored in a transaction table, *D* defined by a set of "SNP" items, $i_1, i_2, \ldots , i_k$. Each item i_k, is associated with a defined genotype. An example is given in the following Table II. Genotype of each attribute is categorical and every individual, X, has exactly the same set of SNPs $\{i_1, i_2, \ldots , i_k\}$ with one of the possible genotype values. Let SNP 1 in Table II, as an example, which has two alleles "A" and "C" and the possible genotypes are "AA",

"AC" and "CC", which "AA" and "CC" are homozygous and "AC" is heterogonous genotype. The total of SNPs in our dataset is around 15 thousand and the sample with diabetes is 1999 records.

B. Ranking Procedure

In selecting features within the SNP datasets, several ranking score formulations applied are information gain [9], [42]. The ranking procedure with information gain and gain ratio defines the order of features to be eliminated.

C. Evaluating Feature Sets Using Classification Methods

The subsets of SNPs are constructed based on n = [500, 1000, 2000, 3000, …] and evaluation of the subsets of datasets is by performing the classification learning using different sets of features. The validation accuracy is used to decide whether the chosen subset of features is permanently eliminated. This process is repeated for each new subset of features and the validation accuracy is estimated based on the built classifiers. If the obtained validation accuracy for the current subset is higher than the accuracy of the previous, then more features need to be included based on their ranking values. The iteration is stopped whenever the validation accuracy of the new subset of features is lower than the one for the previous selected features. Then, the current subset of features is considered as the final subset of features. Otherwise, the procedure is repeated with the lowest ranked features eliminated. The best set of features with the highest validation accuracy is chosen and is called the feature set of best prediction. In terms of selecting the classification method, Support Vector Machine (SVM) is mostly being implemented in SNPs dataset classification [9], [39], [40]. In addition, Decision Tree is also a common method being used [4], [42]. To have a better understanding of this dataset challenges, the study explored Naive Bayes for classification and gain ratio as well as relief measure as feature selectors. In order to perform ranking procedure and classification methods, we set up the experiments using Orange data mining tool box. The script was done in Python and modified based on ranking scores selected and the classification methods of the study.

D. Allelic Test for the Disease and Weight Assignment for Each SNP

The allelic test in genome wide studies is referred to testing a genetic marker for association with a disease in a sample of unrelated subjects. A genetic association case-control study compares the frequency of alleles or genotypes at genetic marker loci, usually SNPs, in individuals from a given population that consists of samples with and without a given disease trait, in order to determine whether a statistical association exists between the disease trait and the genetic marker [2], [4]. The tests are done separately for each individual SNP. The data for each SNP with minor allele *a* and major allele *A* can be represented as a contingency table of counts of disease status by either each genotype count (minor-minor, minor-major and major-major) or allele count (for minor and major).

The allelic Odds Ratio (OR) describes the association between disease and allele by comparing the odds of disease in an individual carrying allele A to the odds of disease in an individual carrying allele a [15], [16]. The genotypic ORs describe the association between disease and genotype by comparing the odds of disease in an individual carrying one genotype to the odds of disease in an individual carrying another genotype. There are two genotypic ORs namely one comparing the odds of disease between individuals carrying genotype A/A and those carrying a/a and the other comparing the odds of disease between individuals carrying genotype a/A and those carrying genotype a/a [16]. Next, the classical Pearson, χ^2 test is also measured to find the significant SNPs to be considered in the itemset mining, which is less than 0.05. Later, the odd ratio value is used to determine the weight of each SNP, following the rules in [43] as shown in Table III below:

TABLE III: ODD RATIO DESCRIPTION

Value of OR	Description
OR=1	Exposure does not affect odds of outcome
OR>1	Exposure associated with higher odds of outcome
OR<1	Exposure associated with lower odds of outcome

E. Implication of Weighted Itemset Mining

A comparison is done to evaluate the itemsets found based on the support and weighted support value. Fig. 2 below shows the flow of the process in performing the experiment for this particular research.

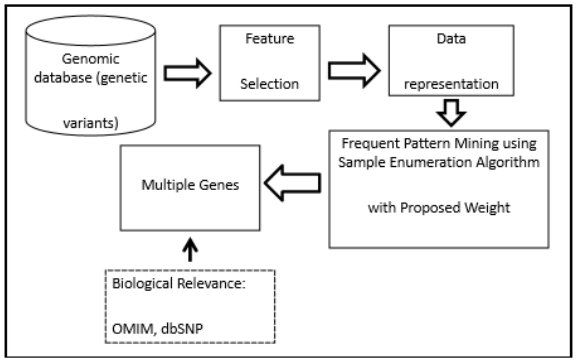

Fig. 2. Experimental flow.

Weight based itemset mining is crucial in that the approach not only reduces search space but also extracts more important patterns. Normally these items may have the presented value, such as products in a shop with its price. However, these real values of items are not suitable for weight values due to big variations and wide ranges, so according to [37], the normalization process is needed to adjust for any differences among the data. In order to determine the weight, a study needs to be done to formulate the calculation. For instance, the final weight of each item is calculated as a basic weight (a) plus the normalized weight of each item. A weight of an item is a non-negative real number which is assigned to reflect the importance of each item in the transaction database [37]. Given a set of items, $I = \{i_1, i_2,...,i_n\}$, the weight of a pattern P $(p_1, p_2,...,p_k)$ is formally defined in Equation (4).

$$\text{weight}(P) = \frac{\sum_{\text{length}(pi)}^{i} \text{weight}(pi)}{\text{length}(pi)} \quad (4)$$

The final weights of items are determined after the normalization process and weights of items are defined within

a specific range. Weights for items are given with $w_{min} \leq w \leq w_{max}$ according to items' importance or priority and the weights are normalized as $min_w \leq w \leq max_w$ and those normalized weights are used in the mining process [37].

IV. Result and Discussion

In our experiment, we have used Intel Core™ i5 machine with 8GB RAM, and made use of Python 2.7 for scripting.

A. Feature Ranking and Classification Results

The study used the hold-out method in learning whereby that 70 percent of data is allocated as the training data, and 30 percent is allocated for the testing data. The validation accuracy for each classifier is computed by the percentage of matches between predicted and actual class. Table IV shows the accuracy of each of these combinations.

TABLE IV: Results of Ranking Scores Based on Info Gain, and Gain Ratio with Classifier's Accuracy for Chromosome 16

Chromosome 16				
Ranking Score	No. of SNPs	SVM	Decision Tree	Naive Bayes
		Accuracy	Accuracy	Accuracy
Information Gain	500	0.8987	**0.8283**	0.8477
	1000	**0.8999**	0.8148	**0.8631**
	2000	0.8841	0.8002	0.8542
	3000	none	0.7542	0.7107
	4000	none	0.7434	0.7007
	5000	0.6062	0.6636	0.6112
Gain ratio	500	0.8983	**0.8832**	**0.8909**
	1000	0.8928	0.8613	0.827
	2000	**0.9041**	0.7845	0.7912
	3000	0.9005	0.7707	0.7756
	4000	0.8889	0.7696	0.7653
	5000	0.8983	0.7395	0.7452

TABLE V: SNP Variants with T2D Risks and the Ranking Score

dbSNP	Chromosome	Gene	Ranking based on Score
rs8050136 Zeggini *et al.* [41] Scott *et al*. [44]	16	Fat Mass and Obesity-associated Gene (FTO)	Information Gain: 72 Gain Ratio: 1934
rs9939609 Wellcome [38]	16	Fat Mass and Obesity-associated Gene (FTO)	Information Gain: 96 Gain Ratio: 2000

The results of experiments present the implication of feature selection technique to identify the best subset of SNPs from over 10 thousands original datasets. The results with the highest accuracy is expected to give more meaningful patterns and knowledge. Based on the table, when the ranking score is Information Gain, 500 SNPs shows highest accuracy with Decision Tree and 1000 SNPs shows highest accuracy with SVM and Naïve Bayes. Meanwhile, if the ranking score is Gain Ratio, 500 SNPs gives highest accuracy with Decision Tree and Naïve Bayes and 2000 SNPs gives highest accuracy with SVM. Biological information and literature would also be useful in making further conclusion. So, for that reason, we made a reference to OMIM [45] which compiles most of the available literature, and a list of SNPs that are valuable in this context is shown in following Table V. Since both identified risky SNPs scores are high, we decided to use the minimum number of SNPs for further process. The reduced SNPs set is captured from Information Gain score with 1000 SNPs.

B. Allelic Test and Weight Assignment

Next, we performed further allelic association test for 1000 SNPs. The SNPs data are measured to get the odds ratio and the significance for the *p* value. For this purpose, a GWAS tool was used, known as the GWASpi_RC_v2.0.2, Genome Wide Association Study pipeline [46]. Figure 3 below shows the control parameters for the data quality.

Samples with missingness >0.5

Markers with missingness >0.05

Markers with Hardy-Weinberg p-Value < (0.05/Markers Nb)

Fig. 3. Control Parameter for Data Quality in GWAS.

Based on the parameters set above, 819 SNPs out of 1000 were selected from Chromosome 16 and these SNPs met the quality control set. Next, the SNPs were filtered based on *p*-value < 0.05, so that only 416 SNPs are listed. Fig. 4 shows the odd ratio for each SNP in Chromosome 16. So, with these 416 SNPs selected (from Chromosome 16), the itemset mining process was executed to find frequent closed itemsets. Meanwhile, in order to apply weighted itemsets, each SNP is assigned with appropriate weight using odd ratio value. In determining the weight, the binning process is applied to determine the normalized weight. The range of odd ratio for 416 SNPs is between 0 and 6.664, and after the binning process is done, 7 bins were applied. The weight is distributed within the range of 0 to 1, as in Table VI below. The decision to determine the weight for each bin (with set of SNPs), is based on the OR effect. If the OR is high, the weight should be higher [44].

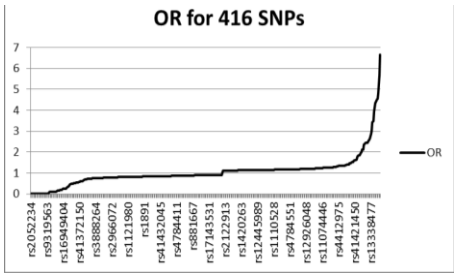

Fig. 4. Odds ratio value for SNP association tests.

TABLE VI: Odds Ratio Value, OR and Weight Range

Bin	OR_min	OR_max	Normalized_weight min	Normalized_weight max	Count of SNPs
1	0	0.351	0	0.1	44
2	0.351	0.702	0.1	0.2	20
3	0.702	1.052	0.2	0.3	161
4	1.052	1.403	0.3	0.4	147
5	1.403	1.754	0.4	0.5	11
6	1.754	2.105	0.5	0.6	6
7	2.105	2.455	0.6	0.7	7
8	2.455	3.858	0.7	0.8	7
9	3.858	5.612	0.8	0.9	6
10	5.612	6.664	0.9	1	2

C. Weighted Frequent Closed Itemsets

However, we firstly have to limit only 100 SNPs out of 416 in this experiment with 522 samples in T2D group from

Chromosome 16. This is due to the huge number of itemsets that were generated if we considered the whole set of SNPs. We run FIM using row enumeration strategies, as each partition file contains 100 SNPs and 29 samples, and row enumeration strategy is more suitable when items are more than the number of samples [25]. After several numbers of experiments, we found that when we run the FIM algorithm with support value at 40, we able to find significant SNPs with T2D, which are rs8050136 and rs9939609. In total, the number of closed itemsets found was 1, 651, 408. However, only 6352 itemsets extracted with identified SNPs. Table VII below shows some of the itemsets with the support value and weighted support values.

TABLE VII: SAMPLE OF SUPPORT VALUE, SUP AND WEIGHTED SUPPORT VALUE, WS

Itemset no.	Sup vs WS	Support	Weighted support	Length of itemsets
7	High sup value, High ws value	48.276	92.564	23
30	Medium sup value, High ws value	44.828	85.952	23
423	High sup value, High ws value	48.276	95.712	23
12	High sup value, High ws value	48.276	89.21	24

TABLE VIII: GENE ID INFO FROM THE ITEMSETS

Number	Gene ID	Number	Gene ID
1.	A2BP1	2.	IL21R
3.	AKTIP	4.	KIAA0556
5.	CARHSP1	6.	KIAA1576
7.	CDH13	8.	LITAF
9.	CNGB1	10.	LOC440389
11.	COTL1	12.	LOC727881
13.	CRISPLD2	14.	LYRM1
15.	DEF8	16.	MT4
17.	FTO	18.	MYH11
19.	GPR56	20.	MYLK3
21.	GRIN2A	22.	PLCG2
23.	GSPT1	24.	PRMT7
25.	HERPUD1	26.	SNX29
27.	HSD17B2	28.	SRCAP
29.	TOX3	30.	TMCO7
31.	ZDHHC7	32.	WWOX

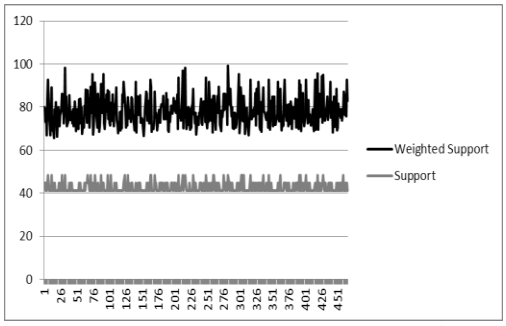

Fig. 5. Weighted support value compared to support value for 465 itemsets only.

Fig. 5 shows the first 465 itemsets with support and weighted support values. Beyond what have been shared above, the gene ID information for each SNP in the extracted SNPs is collected, and in total we have 32 gene IDs from the itemsets, as shown in Table VIII. Common genes extracted in all itemsets in Table VII are HERPUD1, KIAA0556, SRCAP, CDH13, A2BP1, TMCO7, CRISPLD2 and FTO. However, several genes are found separately in different itemsets. LITAF is found in itemset 7 and 12, WWOX is found in itemset 30 and MYH11 is found in itemset 423. The occurrence of each gene was also calculated, and we found that gene FTO, which is associated with obesity gives the highest number of occurrence [42], [45]. Meanwhile, Fig. 6 provides the graphical information on the occurrence of FTO with other gene IDs.

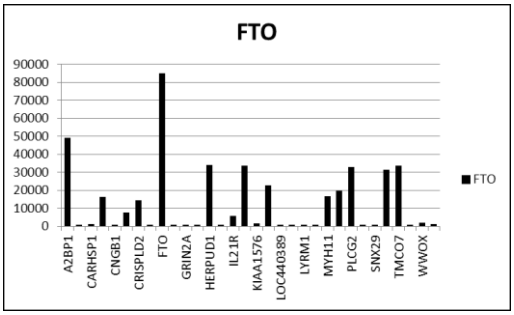

Fig. 6. FTO Occurrence with Other Gene based on SNPs.

We also studied the gene functions in the set of SNPs that we had, as shown in Fig. 7, from Pantherdb [47]. From 32 genes extracted in the SNPs itemsets in Chromosome 16, 25 mapped IDs have its molecular functions with most of the genes related to binding, (CNGB1, A2BP1, PLCG2, GSPT1, LITAF, COTL1, CARHSP1, CDH13).

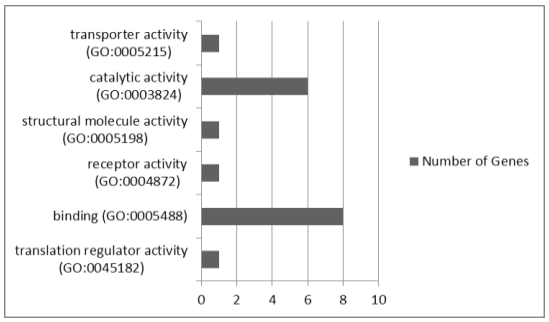

Fig. 7. Known gene function using PANTHER go tool.

V. CONCLUSION

We had implemented filter methods in feature selection and evaluated by accuracy using classifiers. With these methods, we were to find subsets of important SNPs in GWAS that consisted of more than 10 thousand SNPs. In order to perform frequent itemset mining, we calculated the odds ratio and *p*-value and later, limited the study to the most significant SNPs. Then, we proposed a weight assignment for SNPs. The weight assignment was also normalized with the reference to the odds ratio value. Later, the biological information was integrated to the analysis of important SNPs. We found that the weighted itemset mining provided more significant measure for the itemsets. The important and risky SNPs that are meaningful as the disease knowledge, can be extracted with higher weighted support value. Further development and enhancement can contribute in healthcare advancement system.

ACKNOWLEDGMENT

The authors would like to thank to the Research Management Institute, Universiti Teknologi MARA for the REI 19/2015 research grant and Ministry of Education

Malaysia for the research support through FRGS 156/2013.

REFERENCES

[1] B. R. Korf, *Human Genetics and Genomics*, 3rd ed., Blackwell Publishing, 2007.

[2] M. I. McCarthy and J. N. Hirschhorn, "Genome-wide association studies: past, present and future," *Human Molecular Genetics*, vol. 17, pp. R100-R101, 2008.

[3] S. Szymczak, J. M. Biernacka, H. J. Cordell, O. Gonzâlez-Recio, I. R. König, H. Zhang, and Y. V. Sun, "Machine learning in genome-wide association studies," *Genetic Epidemiology*, vol. 33, pp. S51-S57, 2009.

[4] H. J. Cordell, "Detecting gene-gene interactions that underlie human diseases," *Nat Rev Genet*, vol. 10, pp. 392-404, 2009.

[5] G. Atluri, R. Gupta, G. Fang, G. Pandey, M. Steinbach, and V. Kumar, "Association analysis techniques for bioinformatics problems," in *Bioinformatics and Computational Biology*, S. Rajasekaran, Ed., Springer Berlin / Heidelberg, 2009, vol. 5462, pp. 1-13.

[6] J. H. Moore, F. W. Asselbergs, and S. M. Williams, "Bioinformatics challenges for genome-wide association studies," *Bioinformatics*, vol. 26, pp. 445-455, February 15, 2010.

[7] A. DeWan, M. Liu, S. Hartman *et al*., "HTRA1 promoter polymorphism in wet age-related macular degeneration," *Science*, vol. 314, pp. 989-992, November 10, 2006.

[8] R. J. Klein, C. Zeiss, E. Y. Chew, J.-Y. Tsai *et al*., "Complement Factor H Polymorphism in Age-Related Macular Degeneration," *Science*, vol. 308, pp. 385-389, April 15, 2005.

[9] M. Waddell, D. Page, F. Zhan, B. Barlogie, and J. Shaughnessy, Jr., "Predicting cancer susceptibility from single-nucleotide polymorphism data: A case study in multiple myeloma," *Proceedings of BIOKDD '05*, Chicago, Illinois, August 2005, Aug 2005.

[10] A. Kelemen, A. V. Vasilakos, and L. Yulan, "Computational intelligence in bioinformatics: snp/haplotype data in genetic association study for common diseases," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, pp. 841-847, 2009.

[11] A. Calabria, E. Mosca, F. Viti, I. Merelli, and L. Milanesi, "SNPRanker: A tool for identification and scoring of SNPs associated to target genes," *Journal of Integrative Bioinformatics*, vol. 7, 2010.

[12] J. Y. Dai, M. Leblanc, N. L. Smith, B. Psaty, and C. Kooperberg, "SHARE: An adaptive algorithm to select the most informative set of SNPs for candidate genetic association," *Biostatistics*, vol. 10, pp. 680–693, 2009.

[13] Q. He and D-Y. Lin. A variable selection method for genome-wide association studies. *Bioinformatics* , vol. 27, 1–8, 2011.

[14] L. Jie, B. Elizabeth, and P. David, "Predicting breast cancer and prostate cancer susceptibility from single nucleotide polymorphisms," in *Proc. ICML 2013 Workshop on Role of Machine Learning in Transforming Healthcare*, 2013.

[15] D. J Balding, "A tutorial on statistical methods for population association studies," *Nature Reviews Genetics*, vol. 7, no. 10, pp. 781–791, 2006.

[16] H. J. Cordell, "Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans," *Human Molecular Genetics*, vol. 11, no. 20, pp. 2463–2468, 2002.

[17] D. B. Kell, "Genotype–phenotype mapping: Genes as computer programs," *TRENDS in Genetics*, vol. 18, pp. 555–559, 2002.

[18] N. Batnyam, A. Gantulga, and S. Oh, "An efficient classification for single nucleotide polymorphism (snp) dataset," *Computer and Information Science*, Springer, pp. 171-185, 2013.

[19] R. Alzubi, N. Ramzan and H. Alzoubi, "Hybrid feature selection method for autism spectrum disorder SNPs," in *Proc. IEEE Conference on.Computational Intelligence in Bioinformatics and Computational Biology*, 2017.

[20] R. Agrawal, T. Imielinski, and A. Swani, "Mining association rules between sets of items in large databases," in *Proc. ACM SIGMOD Conference*, 1993.

[21] G. Fang, G. Pandey, W. Wang, M. Gupta, M. Steinbach, and V. Kumar, "Mining low-support discriminative patterns from dense and high-dimensional data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 2, pp. 279-294, 2012.

[22] T. Müller, J. Schiffner, H. Schwender, G. Szepannek, C. Weihs, and K. Ickstadt, "Local analysis of SNP data," *Classification as a Tool for Research*, pp. 473-480, Springer Berlin Heidelberg, 2010.

[23] S. H. Park, J. Y. Lee and S. Kim, "A methodology for multivariate phenotype-based genome-wide association studies to mine pleiotropic genes," *BMC Systems Biology*, vol. 5(Suppl 2), S13, 2011.

[24] C. Borgelt, X. Yang, R. Nogales-Cadenas, P. Carmona-Saez, and A. Pascual-Montano, "Finding closed frequent item sets by intersecting transactions," presented at the 14th International Conference on Extending Database Technology, Uppsala, Sweden, 2011.

[25] F. Pan, G. Cong, A. K. H. Tung, J. Yang, and M. J. Zaki, "Carpenter: finding closed patterns in long biological datasets," presented at the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, D.C., 2003.

[26] H. Liu, X. Wang, J. He, J. Han, D. Xin, and Z. Shao, "Top-down mining of frequent closed patterns from very high dimensional data," *Information Sciences*, vol. 179, no. 7, pp. 899-924, 2009.

[27] J. Han, J. Pei, Y. Yin and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," *Data Mining and Knowledge Discovery*, vol. 8, no. 1, pp. 53-87, 2004.

[28] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Discovering frequent closed itemsets for association rules," in *Proc. 7th Int. Conf. Database Theory*, Jerusalem, Israel, 1999, pp. 398–416.

[29] J. Pei, J. Han, and R. Mao, "CLOSET: An efficient algorithm for mining frequent closed itemsets," in *Proc. 2000 ACM-SIGMOD Int. Workshop Data Mining and Knowledge Discovery*, Dallas, TX, 2000, pp. 11–20.

[30] J. Wang, J. Han, Y. Lu, and P. Tzvetkov, "TFP: An efficient algorithm for mining top-K frequent closed itemsets," *IEEE Trans. Knowledge and Data Engineering*, vol. 17, no. 5, pp. 652-653, May 2005.

[31] J. Pei, J. Han and W. Wang, "Constraint-based sequential pattern mining in large databases," in *Proc. the 2002 International Conference on Information and Knowledge Management (CIKM'02*), McLean, VA, 2002, pp. 18–25.

[32] F. Bonchi and C. Lucchese, "On closed constrained frequent pattern mining," in *Proc. the 2004 International Conference on Data Mining*, Brighton, UK, 2004, pp. 35–42.

[33] B. Lucchese, S. Orlando and R. Perego, "Fast and Memory Efficient Mining of Frequent Closed Itemsets," *IEEE Transaction on Knowledge and Data Engineering*, vol. 18(1), 21–36, 2006.

[34] C. H. Cai, A. W. Fu, C. H. Cheng and W. W. Kwong, "Mining Association Rules with Weighted Items," in *Proc. International Database Engineering and Applications Symposium (IDEAS 1998)*, 1998, pp. 68–77.

[35] F. Tao, F. Murtagh and M. Farid, "Weighted Association Rule Mining using Weighted Support and Significance Framework. In: SIGKDD 2003, pp. 661–666, 2003.

[36] C. Wang, W. Wang, J. Pei, Y. Zhu, and B. Shi, "Scalable mining of large disk-base graph databases," in *Proc. the 2004 ACM SIGKDD International Conference on knowledge Discovery in Databases (KDD'04),* Seattle, WA, 2004, pp. 316–325.

[37] U. Yun and J. Leggett, "Wfim: weighted frequent itemset mining with a weight range and a minimum weight," in *Proc. the 2005 SIAM International Conference on Data Mining (SDM'05)*, Newport Beach, CA, 2005, pp. 636–640.

[38] WTCCC, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, pp. 661-678, 2007.

[39] H.-J. Ban, J. Y. Heo, K-S. Oh, and K.-J. Park, "Identification of type 2 diabetes-associated combination of SNPs using support vector machine," *BMC Genetics*, vol. 11, no. 26, 2010.

[40] Z. Wei, K. Wang, H. Q. Qu *et al*., "From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes," *PLoS Genet5*:e1000678, 2009.

[41] E. Zeggini, M. Weedon *et al*., "Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes," *Science*, vol. 316, pp. 1336–1341, 2007.

[42] M. Hajiloo, B. Damavandi, M. HooshSadat, F. Sangi, J. R. Mackey, C. E. Cass, R. Greiner, and S. Damaraju, "Breast cancer prediction using genome wide single nucleotide polymorphism data," *BMC Bioinformatics*, vol. 14, no. Suppl 13, p. S3, 2013.

[43] M. J. Szumilas, *Can Acad Child Adolesc Psychiatry*, vol. 19, no. 3, pp. 227–229, 2010.

[44] L. Scott, K. Mohlke *et al*., "A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants," *Science*, vol. 316, pp. 1341–1345, 2007.

[45] V. McKusick, "Mendelian inheritance in man and its online version, OMIM," *American Journal of Human Genetics*, vol. 80, pp. 588–604, 2007.

[46] F. I. Muñiz-Fernandez, A. Carreño-Torres, C. Morcillo-Suarez, and A. Navarro, "Genome-wide association studies pipeline (GWASpi): A desktop application for genome-wide SNP analysis and management," *Bioinformatics*, 2011.

[47] H. Mi, X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang, and P. D. Thomas, "PANTHER version 11: Expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements," *Nucl. Acids Res*, 2016.

Sofianita Mutalib is in pursuit her doctoral degree in the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia. She had received bachelor and master degree from Universiti Kebangsaan Malaysia in 1998 and 1999. She is actively doing research in applied data mining and data analytics, for various area and different types of data. Her interest also has been shown through yearly publications in proceedings and also journals.

Azlinah Mohamed received her bachelor's degree in Computer Science from Universiti Kebangsaan Malaysia in 1990, M.Sc in Artificial Intelligence from Bristol, UK in 1992 and PhD in Science and System Management from Universiti Kebangsaan Malaysia in 2001. She is currently working as a professor and dean at the Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA, Malaysia. Her research interest includes artificial intelligence, decision support systems, soft computing, and big data analytics.

Shuzlina Abdul Rahman received her bachelor's degree in computer science from Universiti Sains Malaysia in 1996, M.Sc in information technology from Universiti Utara Malaysia in 2000 and PhD in computer science from Universiti Kebangsaan Malaysia in 2012. She is currently working as an associate professor at the Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA, Malaysia. Her research interest includes computational intelligence, machine learning and data analytics & optimization.

Norlaila Mustafa is a senior endrocrinologist who currently works in Hospital Canselor Tuanku Muhriz, Universiti Kebangsaan Malaysia. She has contributed a lot of publications and get involved in research mainly in Type 2 Diabetes. She is one of the Board of Trustees of National Diabetes Institute known as NADI.