An Improved FP-Growth with Hybrid MGSO-IRVM Classifier Approach used for Type-2 Diabetes Mellitus Diagnosis

K.Vembandasamy [#1] T.Karthikeyan[*2]

[#] Computer science & Applications Department,
PSG College of Arts and Science, Coimbatore, Tamilnadu, India
[*]Computer Science Department,
P.S.G. College of Arts and Science, Coimbatore, Tamilnadu, India
[1] vembupsgphd@gmail.com
[2]t.karthikeyan.gasc@gmail.com

Abstract—Diabetes is a chronic disease and major problem of morbidity and mortality in developing countries. Type-2 diabetes mellitus (T2DM) is the most common type of diabetes and accounts for 90-95% of all diabetes. In the medical field a predictive data mining association algorithms is used to diagnose the disease at the earlier stage which helps the physicians in the treatment planning procedure. Recently, an improved Frequent Pattern Growth (IFP-Growth) with Hybrid Enhanced Artificial Bee Colony-Advanced Kernel Support Vector Machine (EABC-AKSVM) classification is introduced with the capability to reduce the number of rules in diagnosing the diabetes. However, the detection accuracy and robust is less. To resolve this problem, an Improved Frequent Pattern Growth (IFP-Growth) with Hybrid Modified Glowworm Swarm Optimization-Improved Relevance Vector Machine (MGSO-IRVM) Classification based Association Rule Mining (ARM) system is proposed in this work to generate effective rules. The proposed work consists of two phases: In first phase, improved FP-growth is proposed to efficiently mine frequent patterns even from uncertain medical database. This is achieved by creating additional array for each item in uncertain transactional database to keep the information of its super-item sets in IFP-tree redundant node generation through this step the computational cost is greatly reduced. In next phase, Hybrid MGSO-IRVM classifier is used to generate the association rules based on the frequent item sets, which avoids rule redundancy and conflicts during the rule mining process. Experimental results show that the proposed model is suitable and alternative model for medical classification to achieve greater accuracy, and to improve medical diagnosis.

Keyword- Association Rule Mining, Frequent Patterns, Glowworm Swarm Optimization, Improved Frequent Pattern Growth

I. INTRODUCTION

Diabetes is one of the most dangerous diseases that causes of death [1]. Diabetes is metabolic disorder that occurs due to failure of body due to produce insulin properly. According to W.H.O, by 2015 a total of 3 hundred millions of the world population will be affected by diabetes [2]. It has been noticed that diabetes affected a more fatal persons and also the women than men. The cause of worst affect on women is their lower survival rate and poor quality of life. A cause of diabetes is also that many of the peoples don't have knowledge this disease [3]. Human body needs energy for activation the carbohydrates are broken down to glucose. That is the important energy source for body cells. Insulin is necessary to translate the glucose into body cells. The blood glucose is supplied with insulin [4]. In the world, there are many systems that are used for the advanced complication predictions of diabetes symptoms and produce the results on the basis of these symptoms. Most of these systems predict the results based on datasets available in clinical labs. But some the systems predict the causes of diabetes based on the risk factors. Such as Insulin Resistance, age, central obesity, Stress, Polyuria, Polydipsia disease etc, but still the major problem of these systems are to diagnose the disease correctly and costly medical tests [1]. Many data mining techniques are used to solve these problems. Data mining techniques helps the experts and patients to calculate the diabetes risks, on the basis of risks they can know in advance either they affected with diabetes or not.

Data mining is a kind of self-knowledge discovery and a method for the purpose of the analysis of large databases, providing unidentified, secret, meaningful, and functional patterns automatically obtained from largescale databases [6]. Their algorithms like neural networks [6], Support Vector Machines (SVM) [5], decision tree [7] and etc., have been utilized productively in several medical fields. These algorithms have been capable of providing a decent solution in several diseases' diagnostic systems like diabetes [5], heart disease [6], breast cancer [9], etc. In addition, exploitation of data mining schemes makes new information and relationships embedded in large and complex datasets evident through inference and learning new patterns and associations [7,

8]. The utilization of large quantities of patients' data to diagnose disease, by means of data mining schemes, improves the accuracy of these schemes [7, 8].

Farahmandian et al. [5] provided the diagnosis of diabetes on 768 samples with the help of Pima Indians Dataset [10]. Here, the following schemes are used; SVM, KNN, Naive Bayes, ID3, C4.5, C5.0, and CART, 80% of data is used for the purpose of education and 20% for the purpose of testing.

Here, Multi-Layer Perceptron (MLP) neural network is used for the purpose of prediction. MLP has an input layer with eight Pima Indians Dataset features and an output layer that predicts, together with the ninth property which is zero (Normal) and one (Sick). 60% of data is taken for education, 20% is taken for testing data, and 20% for the purpose of application set. The accuracy obtained for education is 97.61% in MLP. In case of [11], using Pima Indians Dataset, the researchers made an attempt to diagnose diabetes. Here, applied the data mining schemes like KNN, Amalgam KNN, K-Means, EM, and ANFIS. Comparing these schemes, amalgam KNN was more accurate than others.

In this paper, Hybrid Modified Glowworm Swarm Optimization- Improved Relevance Vector Machine (MGSO-IRVM) proposed in the data mining action to access information from actual data of patient medical records. It presents a decision-making support through association rule mining based classification with the frequent itemset result of Improved FP-Growth. The rest of the work is as follows: This paper is organized as follow: a brief explained of medical data used and pre-processing is provided in section II. The detailed information is given for each subsection with the proposed work with subsections. Section III gives result and discussion, and finally, in section IV discusses the conclusion with summarization of the result by emphasizing this study and also mentioning for future research.

II. MATERIALS AND METHODS

In this section, the proposed method of Improved FP-Growth with hybrid MGSO-IRVM is explained. And the data preprocessing, association rule mining also discussed in detail.

A. System overview

The Fig.1 illustrates the overall block diagram of the proposed system. Initially, the Type-2 DM patient data is given as input. This may contain unwanted and empty data this significantly reduce the detection accuracy. In next step, preprocessing is done to remove the noisy data. This is done through the Normalization without outliers and Chi Merge discretization method. After that IFP-growth algorithm utilizes for finding frequent patterns of the input dataset. The resultant frequent itemsets from IFP-Growth is then applied to Hybrid MGSO-IRVM Classification algorithm, as a result the accurate result of association rules are generated. The detailed description is explained in further sections.

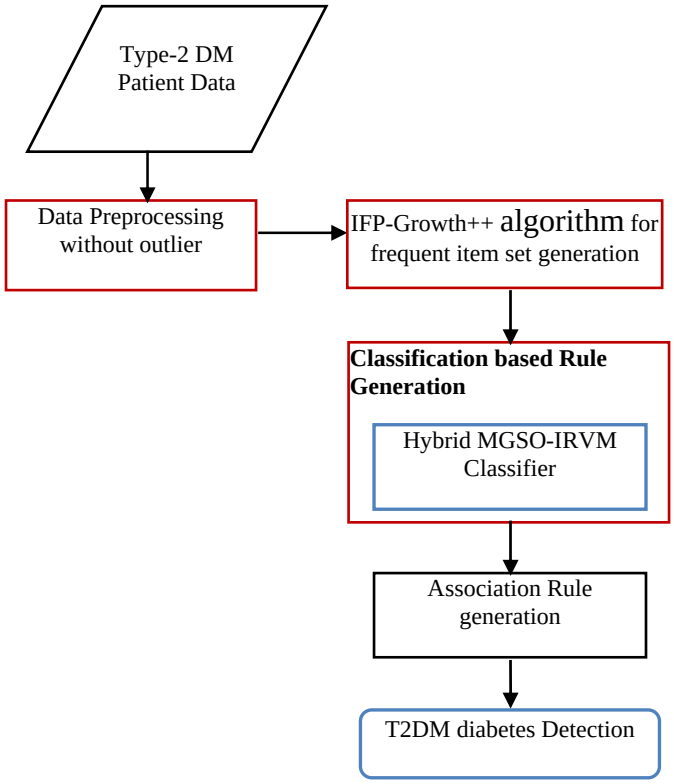

Fig 1: Overall block diagram of proposed method

B. Dataset Pre-processing

The PIMA Indian diabetes dataset was donated by Vincent Sigillito and consists of 768 instances collected. In particular, all patients here are females at least 21 years old of Pima Indian heritage. This is an extremely important step because it can influence the result of a classification algorithm. This module computes tuples with missing values by means of different choices like maximum, minimum, constant, average and standard deviation, prior to applying the normalization scheme. This process provides the treatment of missing value data and subsequently it applies to the second part of data preparation. This step works in two stages. At some point in the first stage, after preprocessing the dataset, work out and discard 5-95% data from the dataset. Accumulate and normalize these discarded data independently which is regarded as outliers.

Normalization is the scaling of data transformation of instances. In a dataset, the instance may have different values ranging from minimum to maximum values. This normalizes all the numerical values present in the dataset. The result values occupy the range between [0, 1]. Here the data normalization is considered between standard deviation and mean square error. It improves the robustness between two data's relationship.

$$v' = \frac{\sum_{i=1}^{n} v - \overline{A}_n}{\sum_{i=1}^{n} \sigma A_n} \quad (1)$$

$\overline{A}_n$ and σA_n stands for the mean, standard deviation and value of attribute $A = \{a_1, a_2, \ldots a_n\}$. In this work the attribute *A* value is 13. Then the discretization is the process of converting the quantitative data into qualitative data. Quantitative data are more commonly presented in many data mining applications. But learning algorithms usually considers qualitative data; with a quantitative data the learning algorithm behaves less efficient and less effective values. Chi Merge is utilized as a discretization method [12]. The speed of proposed algorithm can be improved by using discretized variables.

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}, O_{ij} \quad (2)$$

Where χ^2 is the observed frequency of interval i for class j and e_{ij} is the expected frequency $(R_i * C_j) / N$

C. Frequent Itemset Generation

The proposed IFP-Tree consists of mainly two elements- the tree and a table [13]. The tree represents the correlation among the items more specifically and table is used to store the spare items. It is called as spare table (Stable) which has two columns as item_name and frequency. Item_name is the name of the items and frequency means how many times it occurs in Stable. The main reason to introduce the spare table is, in traditional FPtree a lot of branches are created and the same item appears in more than one node. But in the proposed IFP-tree every distinct item has only one node. So it is simpler and efficient for further processing. The overall process of frequent itemset generation is given in Algorithm1.

1)Lemma 1: Let A is the set of items. Spare item S is defines as $S \subset A$. There are two cases where an item is considered as spare.

Case 1: When the item has no single edge from the current root to its node. That means the item is already exists in the IFP-tree (Improved FP-Tree). In proposed IFP-tree, the root is not fixed, it changes.

Case 2: The transaction items that do not contain the most frequent item.

2)Lemma 2: The frequency of an item in the IFP-tree is the number of time, it occurred in the IFP- tree. It is presented as {item_name: frequency of IFP-tree}.

3)Lemma 3: An uncertain item is an item $x \in W$ whose presence in a transaction $t \in T$ is defined by an existential probability $P(x \in t) \in (0, 1)$. A certain item is an item where presence of an item x is either 0 or 1.

4)Lemma 4: An uncertain transaction t is a transaction that contains uncertain items. A transaction database containing uncertain transactions is called an uncertain transaction database (D).

Case 1: In the process of constructing the UIFP tree (Uncertain Improved FP), the frequent 1-itemsets with their expected support ($expSup$) larger than or equal to the minimum count were first found. Each transaction in the uncertain database was then updated to keep only the frequent 1-items (if $expSup(I) = n \times s$). The construction process of the UIF tree was a little like building an IFP-tree except that the same items with different existential probabilities was put in different nodes. Only the same items with the same existential probabilities in transactions were merged together in the tree.

5)Algorithm 1:

Find Frequent Item sets using IFP-Growth algorithm

Input: S-Support, C-Confidence, F-Frequency, R-Root, K_1-Frquent-1-itemset
Output: The complete set of frequent patterns

1. Procedure call Mining Frequent Item set (IFP-tree, S, C, F, R, j, n, K_1, $expSup$)
2. Sort F in support descending order as i, the list of frequent items.
3. Create the root of the tree R.
4. For each transaction in transaction database **do**
5. Let a descending ordered transaction is represented by $[i|Q]$.where p is the first item and Q is rest of the items in transaction.
6. If i= the most frequent item, do the following step 9 and step 12.Otherwise go to step 15.
7. for each item i in FP-tree where $(i! = R)$ **do**
8. **if** $i.S = i.F$ **then**
9. frequency of the frequent item set, $K = i.F$
10. Generate item set, $P = (i\mu)$ with the frequency value of the tree.

/* μ = All possible combinations of the item and nodes with higher frequency in FP-tree */
Frequent item set is written in $\{P: K\}$ format.

11. **else if** $i.S > i.F$ **then**
12. Frequency of frequent item set $K = i.F + C$

* C=frequency in Stable count*/

13. Generate item set, $P = (ia)$

/*a = All possible combination of item and all intermediate nodes up to most frequent item node in IFP-tree*/

14. **else** Frequency of the frequent item set $K = i.F$
15. Generate item set, $P = (i\beta)$
16. **for** each number of transactions $n \geq ++j$ do
17. **for** each frequent item $i \in K_1$do
18. **if** $expSup(i) = n \times s$
19. **then** i is an uncertain frequent-1 item sets
20. **end for**
21. **end for**

Now, it is notable that in every previous approach to find the frequent item set generating like [14] generates the 1-itemset, 2-itemset and so on. But in this approach all types of possible item set which satisfy the minimum user defined support is generated. This is make the proposed algorithm efficient than the others. Now it is quite easy to define the association rules from the frequent item set. By creating additional array (expAry) for each item in uncertain transactional database to keep the information of its super-item sets in CUIFP-mine, redundant node generation and the computational cost in the mining process is greatly reduced.

D. The Rule Generation Stage

The frequent itemsets produced are used to generate association rules that satisfy minimum support and minimum confidence. Generally, in association rule mining, any item that passes minsupp is known as a frequent itemset. The frequent itemsets are generated using IFP-Growth explained in above section. When the frequent items have been discovered, classification based on association rules algorithms extract a complete set of Class Association Rules (CAR) for those frequent items that pass minconf. The key operation of CAR-RG (Class Association Rule-Rule Generation) is to discover all ruleitems that have support beyond minsup. A ruleitem is of the form: $< \mathrm{I}, \mathrm{y} >$ where I stands for a set of items, $y \in Y$ indicates a class label. The support count of the $\mathbb{I}$ (called Isupcount) stands for the quantity of cases in $\mathbb{D}$ that contain the $\mathbb{I}$. The support count of the ruleitem (called rulesupCount) stands for the number of cases in D that contain the I and are labeled with class y. Every ruleitem fundamentally represents a rule: $\mathrm{I} \rightarrow \mathrm{y}$, whose support is $(\mathrm{rulesupCount}/|\mathrm{D}|) * 100\%$, where $|\mathrm{D}|$ stands for the size of the dataset, and whose confidence is $(\mathrm{rulesupCount}/\mathrm{IsupCount}) * 100\%$. Ruleitems that meet the minsup condition are called frequent ruleitems, at the same time the rest are called infrequent ruleitems.

For instance, let the ruleitem is given as: $<\{(A,1),(B,1)\},(class,1)>$, where *A* and *B* are attributes. When the support count of the $I\{(A,1),(B,1)\}$ is 3, the support count of the ruleitem is 2, and the overall number of cases in *D* is 10, subsequently the support of the ruleitem is 20%, and the confidence is 66.7%. When minsup is 10%, subsequently the ruleitem satisfies the minsup criterion, can say it is frequent. For the entire ruleitems that have the same I, the ruleitem with the maximum confidence is selected as the Possible Rule (PR) representing this set of ruleitems. When there are more than one ruleitem with the same maximum confidence, at that time randomly choose one ruleitem. For instance, consider two ruleitems that comprise the same : $1.<\{(A,1),(B,1)\},(class:1)>$ $2.<\{(A,1),(B,1)\},(class:2)>$. Consider the support count of 1 is 3. The support count of the initial ruleitem is 2, and the second ruleitem is 1. Subsequently, the confidence of ruleitem 1 is 66.7%, at the same time the confidence of ruleitem 2 is 33.3% with these two ruleitems, only generate one PR $(assume\ |D|=10)$: $(A,1),(B,1)\rightarrow(class,1)[supt=20\%,confd=66.7\%]$. When the confidence is more than minconf, can confirm the rule is accurate. The set of Class Association Rules (CARs), thus consists of the entire PRs that are both frequent and accurate. To achieve this process, a Hybrid MGSO-IRVM classifier with high accurate results is proposed.

E. Improved Relevance Vector Machine

The given training inputs$\{x_i, t_i\}_{i=1}^{n}$, where $x_i \in R^n, t_i \in \{0,1\}$ and n is the number of samples. The RVM makes predictions for new inputs $\hat{x}$ based on the SVM-like function; the model takes the form of a linear combination of basic functions transformed by a logistic sigmoid function

$$y(\hat{x},w)=\sigma\left(\sum_{i=1}^{n}\omega_i k(x_i,\hat{x})\right)=\sigma(w^T K) \tag{3}$$

Where $k(\hat{x})=[k(x_1,\hat{x})\ldots k(x_n,\hat{x})]^T$ is the kernel function vector, $w=(\omega_1\ldots\omega_n)^T$ is the weight vector, and $\sigma(.)$ is the logistic sigmoid function defined by:

$$\sigma(a)=\frac{1}{1+\exp(-a)} \tag{4}$$

The logistic sigmoid function satisfies the following symmetry property:

$$\sigma(-a)=1-\sigma(a) \tag{5}$$

Therefore, RVM model can be used as the posterior probability. For the input $\hat{x}$, the posterior probability of class c_1 can be written as:

$$p(t=1|\hat{x})=y(\hat{x},w) \tag{6}$$

Similarly, the posterior probability of class c_2 can be written as:

$$p(t=0|\hat{x})=1-y(\hat{x},w) \tag{7}$$

Because its model can be treated as the posterior probability, RVM adopts a Bayesian probabilistic framework to train the model. The key feature of RVM is using the Automatic Relevance Determination (ARD) prior over the weight vector **w,** in which there is a separate hyper parameter α_i for each of the weight parameters ω_i. During the inference procedure, many of the hyper parameters are driven to large values, so that the corresponding weights are effectively forced to zero. Thus the corresponding kernel functions can be pruned out, resulting in a sparse model. The inputs x_i corresponding to the remaining nonzero weights are called relevance vectors.

For an input vector $\hat{x}$, the RVM decision model, as pre-defined by Equation(3), can be rewritten only based on the w_{MP} and RVs as follows

$$y(\hat{x},w_{MP})=\sigma\left(\sum_{x_i\in RVS}\omega_i k(x_i,\hat{x})+\omega_0\right) \tag{8}$$

As can be seen in Equations (3) and (8), kernel function plays an important role in the RVM decision model. There are several common kernel functions for selection, such as linear, polynomial, sigmoid, Gaussian radial basis function (RBF) and so on. In this improved RVM the Elliptical Radial Basis Function (ERBF) used for kernel function.

$$(x,z)=\exp\left(-\sum_{i=1}^{D}(x_i-z_i)^2/(\sigma_i^2.r^2)\right) \tag{9}$$

Where x and z are D-dimension feature vectors (i.e. $=(x_1,\ldots,x_D)^T, z=(z_1,\ldots,z_D)^T$), r is scale factor, σ_i^2 variance. Modified glowworm swarm optimization approach is proposed for optimization of the hyper parameter for RVM.

F. Modified Glowworm Swarm Optimization

Entire process of GSO algorithm includes four steps: deployment of glowworms phase, luciferin-update phase, movement phase and local-decision domain update phase. Deployment of glowworms phase: in the phase, the purpose is to enable the glowworms to be dispersed in the entire objective space. Each glowworm contains equal quantity of luciferin and sensor range. Luciferin-update phase: during the luciferin update phase, each glowworm changes luciferin value according to the objective function value of their current location. The luciferin update rule is given by:

$$l_i(t+1) = (1-\rho)l_i(t) + \gamma j_i(t+1) \tag{10}$$

Where $\rho(0 < \rho < 1)$ is the luciferin decay constant,$j_i(t)$ is the luciferin enhancement constant and J (t) i indicates the objective function value at glowworm i' s location at time t.

Movement phase: during the movement phase, each glowworm selects a neighbor that has higher luciferin value and moves toward it using a probabilistic mechanism. The probability of glowworms $p_j(t)$ moving towards a neighbor j is based on the Eq. (10) at iteration t:

$$p_j(t) = \frac{(l_j(t) - l_i(t))}{\sum_{k\epsilon n_i(T)}(l_k(t) - l_i(t))} \tag{11}$$

Where $l_i(t)$ is the luciferin value of glowworm i , d(i, j) is the Euclidian distance between glowworms i and j . The movement of glowworms i is as follows:

$$x_i(t+1) = x_i(t) + s\left(\frac{x_j(t) - x_i(t)}{\|x_j(t) - x_i(t)\|}\right) \tag{12}$$

Where s is the step-size. Local-decision domain update phase: when the number of neighbor changes, local-decision domain needs update at each of iteration, local-decision domain update rule can be presented by the following equation:

$$r_d^i(t+1) = \min\{r_s, \max\{0, r_d^i(t) + \beta(n_t - |N_i(t)|)\}\} \tag{13}$$

Where $r_d^i(t+1)$ is the local-decision domain of glowworm i at the t+1 iteration, β is an constant parameter that affects the rate of change of the neighbor domain, n_t is a threshold that is used to control the number of neighbors.

In this modified GSO, firstly introduce Tent map of chaos into deployment of glowworms, in which chaos is a universal non-linear phenomenon that has better ergodicity and randomness. At the deployment phase of glowworms, the improved deployment method makes glowworms better randomness in stage, the algorithm has strong ergodicity which enables the glowworms to search the optimal value accurately. Mathematical expression of Tent map as show below:

$$x_{k+1} = \begin{cases} 2x_k, 0 \le x_k \le 0.5 \\ 2(1-x_k), 0.5 \le x_k \le 1 \end{cases} \tag{14}$$

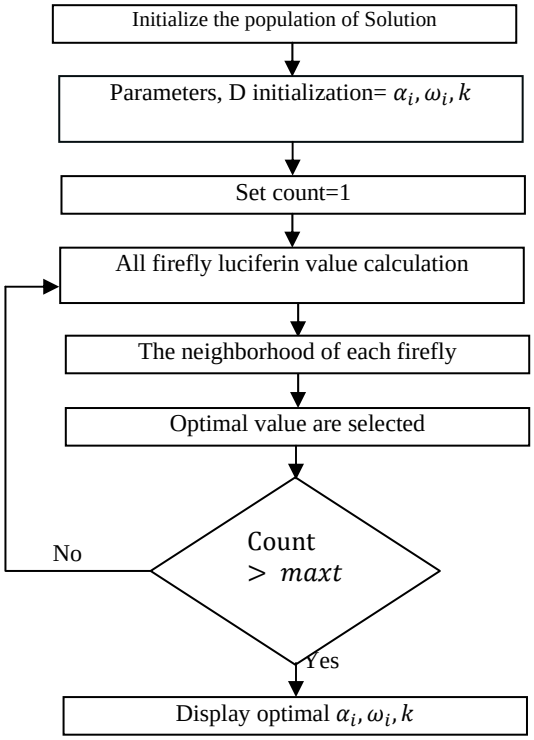

Fig 2: Flow Diagram of MGSO-IRVM

G. Hybrid MGSO-IRVM classifier

In this sub section, CAR-CB (Classification Association Rule-Classifier Bulider) algorithm is presented for building a classifier using CARs. To produce the best classifier with the complete set of rules would involve evaluating all the possible subsets of it on the training data and selecting the subset with the right rule sequence that gives the least number of errors. There are 2^m such subsets, where m is the number of rules, which can be more than 10,000, not to mention different rule sequences. This is clearly infeasible. The GSO operates the firefly luciferin value. Fig.2 shows the simplified form of the proposed prediction model while the flow of Hybrid MGSO algorithm with IRVM. From this Fig, it can be seen that the firefly approach is applied in initialization phase prior to producing new food solution.

Finally, the refined rule generated by IFP-Growth-Hybrid MGSO-IRVM is given below for example, which is used for the prediction of and cause of diabetes. Rule 1.{Urine Alb. <300}{Heart problem is absent}{ creatinine is Negative}{TG < 250}{Uric Acid is Absent}{LDL is Low}->{T2DM is present} 96.66%.

III. Experimental Results And Discussions

The Pima Indian Diabetes Dataset is publicly available at UCIrvine Machine Learning Lab and widely used as a standard for testing the accuracy of diabetes status using data mining algorithms (http://www.uci.edu/). The dataset has 512 training examples and 256 examples as test data with 8 attributes. The attribute are listed in TABLE I. The 13th attribute is a diabetic class which has two values tested positive and tested negative of nominal type. Out of 768 patients 500 are tested negative (Class=0) and 268 are tested positive (Class=1).

TABLE I. Attributes in Pima Indian Diabetes Dataset

Attribute No.	Attribute Description	Type
a_1	PREG Numbers of time pregnant	Numeric
a_2	PGGT Plasma glucose concentration in an oral glucose tolerance test	Numeric
a_3	BP Diastolic blood pressure (mmHg)	Numeric
a_4	SKIN Triceps skin fold thickness (mm)	Numeric
a_5	INS Serum insulin (μU/ml)	Numeric
a_6	MASS (BMI) Body mass (thin, medium, overweight)	Numeric
a_7	PEDI Diabetes pedigree function	Numeric
a_8	AGE Age of patient (years)	Numeric
a_9	Hyperlipidemia (true, false)	-
a_{10}	Fasting blood sugar (FBS) (< 126 mg/dl, ≥ 126 mg/dl)	Numeric
a_{11}	Instant blood sugar (< 200 mg/dl, ≥ 200 mg/dl)	Numeric
a_{12}	Diabetes Gest history (true, false)	-
Y	DIABETES Diabetes diagnose results ("tested_positive", ―tested_negative")	Nominal

After, the preprocessing step discussed in previous section, applied CAR-RG based on Hybrid MGSO-IRVM to produce the refined rules. The results of proposed IFP-Growth++ with Hybrid MGSO-IRVM are compared with existing IFP-Growth++ with Hybrid EABC-AKSVM, CFP-Growth++ with MPSO-LSSVM and Ant Enhanced FP-Growth based classification technique in term of accuracy rate, runtime and number of rules produced etc.

A. Convergence performance

The comparison in terms of convergence between IFP-Growth++ with Hybrid MGSO-IRVM, IFP-Growth++ with Hybrid EABC-AKSVM, CFP-Growth++ with MPSO-LSSVM and Ant Enhanced FP-Growth [15] is illustrated in Fig 3. The efficiency of proposed model can be seen from the narrow span showed in the graph, where the predicted value by IFP-Growth++ with Hybrid MGSO-IRVM is more accurate than IFP-Growth++ with Hybrid EABC-AKSVM, CFP-Growth++ with MPSO-LSSVM and Ant Enhanced FP-Growth. In addition, the proposed model also offers better convergence performance as compared to standard existing algorithms.

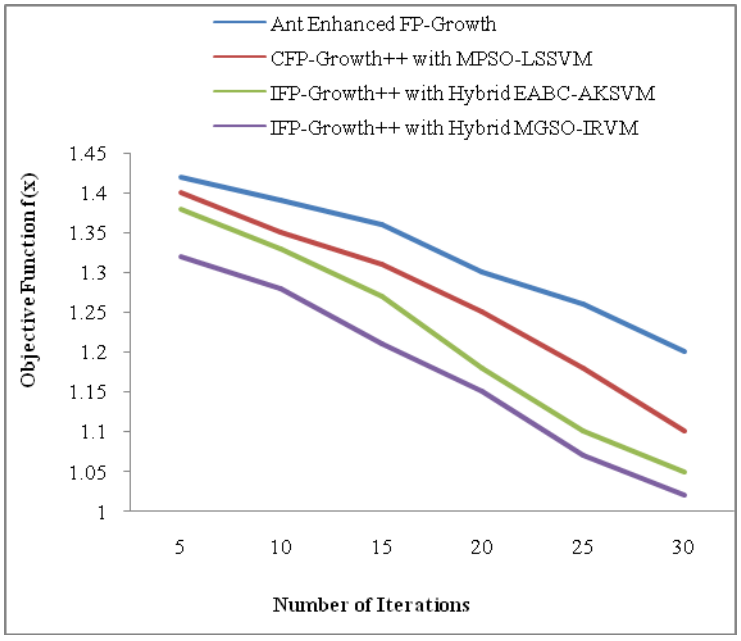

Fig 3: Convergence performance

B. Average number of Identified Rules Comparison

The comparison in terms of average number of rule generation between IFP-Growth++ with Hybrid MGSO-IRVM, IFP-Growth++ with Hybrid EABC-AKSVM, CFP-Growth++ with MPSO-LSSVM and Ant Enhanced FP-Growth is illustrated in Fig.4. When the number of attributes is increases the average number of rules is increases. However, average number of rules generated by IFP-Growth++ with Hybrid MGSO-IRVM is less when compared to existing algorithm. The efficiency of proposed model can be seen from the graph, where the average number of rules predicted IFP-Growth++ with Hybrid MGSO-IRVM is more accurate than existing algorithms.

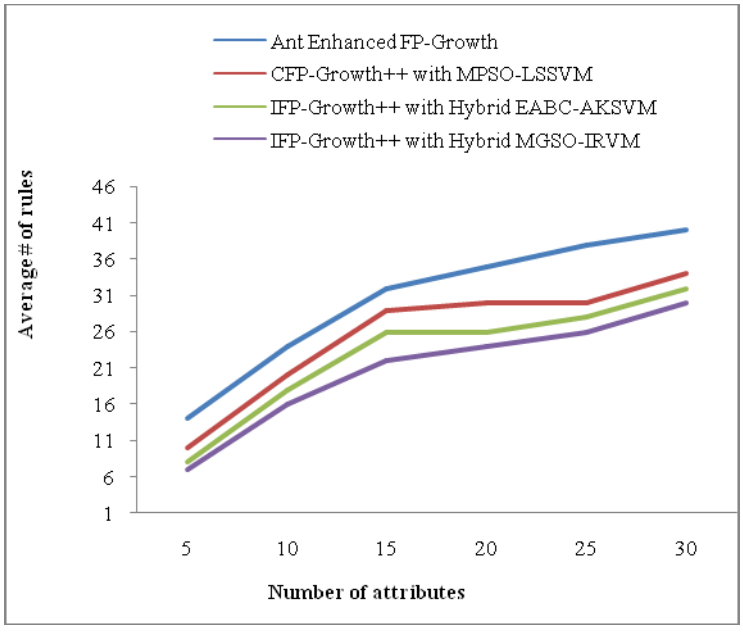

Fig 4: Average number of Identified Rules Comparison Results

C. Processing Time Comparison

The IFP-Growth algorithm discovers frequent item sets and Fig 5 shows much greater efficiency than existing algorithm IFP-Growth++ with Hybrid EABC-AKSVM, CFP-Growth++ with MPSO-LSSVM and Ant Enhanced FP-Growth. When the support value is increases the process time decreases. The proposed algorithm takes less computation time to generate rules when compared to existing system. The algorithm IFP-Growth++ with Hybrid MGSO-IRVM is reportedly working efficiently and in many cases, it's much faster than existing algorithms.

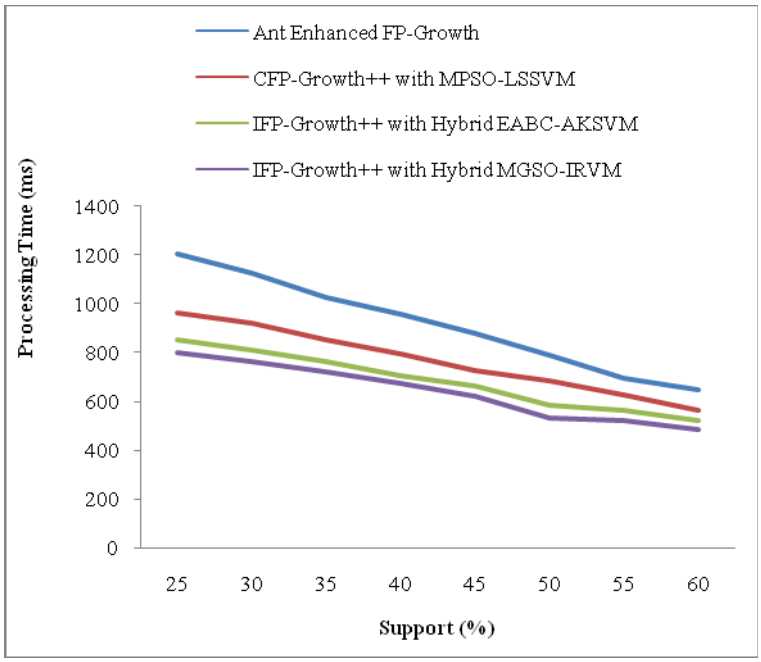

Fig 5: Processing Time Comparison

D. Accuracy Comparison

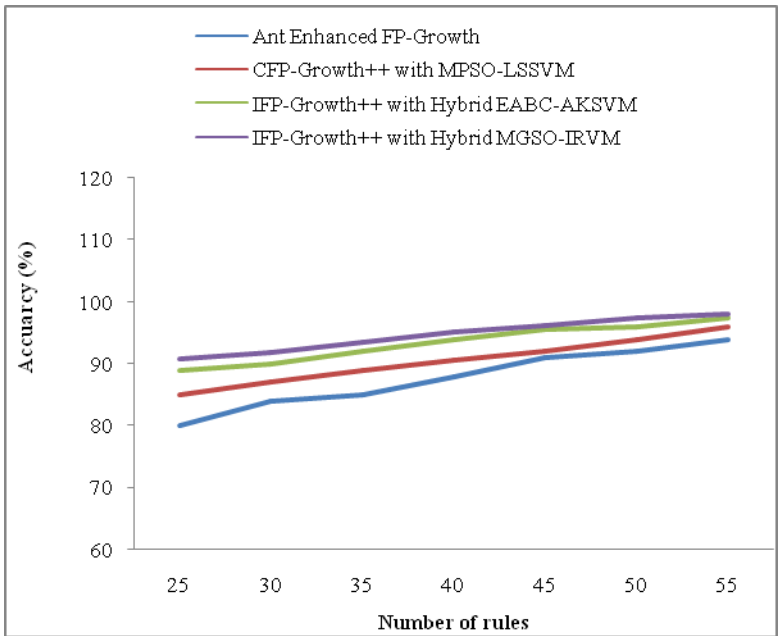

Fig 6: Accuracy Comparison

The IFP-Growth algorithm discovers frequent item sets and Fig 6 shows much greater accuracy results than existing algorithm IFP-Growth++ with Hybrid EABC-AKSVM, CFP-Growth++ with MPSO-LSSVM and Ant Enhanced FP-Growth. When the number of rules generation is increases the accuracy of the result is increases. The proposed algorithm produces high accuracy rate when compared to existing system. The algorithm IFP-Growth++ with Hybrid MGSO-IRVM is working effectively and in many cases produces high accuracy rate than existing algorithms.

IV. CONCLUSION

The prevalence of diabetes is increasing among young adults and old age people. In this paper, IFP-Growth++ with Hybrid MGSO-IRVM is proposed to diagnosis type 2 diabetes mellitus. The proposed method is used preprocessing in order to improve quality of data which includes normalization and data discretization. In later stage medical patient record of type 2 diabetes mellitus (TTD) are collected and analyzed with IFP-Growth and Classification based ARM to extract the information from TTD patient. Based on that formation the frequent itemsets are generated using IFP-Growth with the ability to meet uncertain database. The storage

capacity and computation cost is reduced through this method. Then, the proposed MGSO-IRVM classification based association rule generation refines the optimal association rules. This method is highly faster and more effective in terms of predicting and generating the rules. The experiment has been successfully performed with the Pima indian datasets in which the proposed approach is achieved high accuracy compared than existing data mining techniques. In future, an effective approach is proposed to mine association rule from the database with missing values without generating candidate itemsets and to apply other classification methods.

REFERENCES

[1] A. A.Chaudhari, and S.P.Akarte, "Fuzzy & Datamining based Disease Prediction Using K-NN Algorithm," International Journal of Innovations in Engineering and Technology (IJIET), vol. 3, no. 4,pp.9-14, Apr.2014.

[2] Sumathy, T.Mythili, Praveen Kumar, T. M. Jishnujit, and K.Ranjith Kumar, " Diagnosis of Diabetes Mellitus based on Risk Factors, " International Journal of Computer Applications, vol.10, no.4, Nov. 2010.

[3] Aiswarya Iyer, S. Jeyalatha, and Ronak Sumbaly, "Diagnosis Of Diabetes Using Classification Mining Techniques," International Journal of Data Mining & Knowledge Management Process (IJDKP), vol 5, no.1, pp.1-14, Jan. 2015.

[4] M. Durairaj, and G. Kalaiselvi, "Prediction of Diabetes using Soft Computing Techniques- A Survey" International journal of scientific & technology research, vol. 4, no.03, pp.190-192, Mar. 2015.

[5] M.Farahmandian, Y.Lotfi, and I.Maleki , "Data Mining Algorithms Application in Diabetes Diseases Diagnosis: A Case Study," MAGNT Research Report, 3(1), 989-997, 2015.

[6] F.S. Gharehchopogh, and Z.A. Khalifelu, "Neural Network Application in Diagnosis of Patient: A Case Study", International Conference on Computer Networks and Information Technology (ICCNIT 2011), IEEE, Abbottabad, Pakistan, pp. 245–249, July 2011.

[7] F.S Gharehchopogh, P.mohammadi, and P.Hakimi, "Application of Decision Tree Algorithm for Data Mining in Healthcare Operations: A Case Study," International Journal of Computer Applications (IJCA), vol.52, no. 6, pp. 21-26, Aug.2012.

[8] F.S. Gharehchopogh, "Approach and Review of User Oriented Interactive Data Mining", 4th International Conference on Application of Information and Communication Technologies (AICT2010), Digital Object Identifier, IEEE, Tashkent, Uzbekistan, pp.1-4, Oct.2010.

[9] M. F. B. Othman, and T. M. S. Yau, "Comparison of Different Classification Techniques Using WEKA for Breast Cancer," Kuala Lumpur International Conference on Biomedical Engineering, vol. 15, pp 520-523, 2006.

[10] A .Al-Rofiyee, M. Al-nowiser, N. Al-Mufadi, and M. A. AL-Hagery, "using prediction methods in data mining for diabetes diagnosis", posters, May.2014.

[11] V. Vijayan , and A. Ravikumar, "Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus", International Journal of Computer Applications , vol.95, no.17, pp.12-16, Jun. 2014.

[12] T.Karthikeyan, and Vembandasamy K, "A Novel Algorithm to Diagnosis Type II Diabetes Mellitus Based On Association Rule Mining Using MPSO-LSSVM with Outlier Detection Method", Indian Journal of Science and Technology, vol.8, Apr.2015.

[13] T.Karthikeyan, and K.Vembandasamy, "An Intelligent Type-II Diabetes Mellitus Diagnosis Approach using Improved FP-growth with Hybrid Classifier Based Arm," Research Journal of Applied Sciences, Engineering and Technology, vol.11, no.5, pp.549-558, 2015.

[14] I.Tudor, "Association rule mining as a data mining technique," BULETINUL universitatii Petrol-Gaze din Ploiesti,vol.60, no.1,pp. 49-56, 2008.

[15] T.Karthikeyan, and K. Vembandasamy, "A refined continuous ant colony optimization based FP-growth association rule technique on type 2 diabetes,"IRECOS, vol.9, no.8, pp.1476–83, 2014.

AUTHOR PROFILE

Prof.K.Vembandasamy received his MCA from Bharathiar University. He has a teaching experience of eight years. Currently he is doing Phd in computer science from Bharathiar University His research interests are data mining and web mining. Presently he is working as an Assistant Professor in Computer science & Applications Department, of PSG College of Arts and Science, Coimbatore; he has published one national and two international papers. He has actively participated and organized national and international conferences.

Prof. Thirunavukarasu Karthikeyan received his graduate degree in Mathematics from Madras University in 1982. He received his Post graduate degree in Applied Mathematics from Bharathidasan University in 1984. He has completed his doctorate in Computer Science from Bharathiyar University in 2009. Presently he is working as an Associate Professor in Computer Science Department of P.S.G. College of Arts and Science, Coimbatore. His research interests are Image Coding, Medical Image Processing, Data Mining and Software Engineering. He has published many papers in national and international conferences and journals. He has completed many funded projects with excellent comments. He has contributed as a program committee member for a number of international conferences. He is the review board member of various reputed journals. He is board of studies member for various autonomous institutions and universities.