Research Journal of Applied Sciences, Engineering and Technology 11(5): 549-558, 2015
DOI: 10.19026/rjaset.11.1860
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: May 16, 2015 Accepted: June 22, 2015 Published: October 15, 2015

Research Article

An Intelligent Type-II Diabetes Mellitus Diagnosis Approach using Improved FP-growth with Hybrid Classifier Based Arm

[1]T. Karthikeyan, [1]K. Vembandasamy and [2]B. Raghavan
[1]PSG College of Arts and Science, Coimbatore, India
[2]Department of Biochemistry, PSG College of Arts and Science, Coimbatore, India

Abstract: Diabetes mellitus has turned out to be a common chronic disease that affects between 2 and 4% of the total population. Recently, most of the system uses association rule mining for diagnosing type-II diabetes mellitus. The most vital concern of association rules is that rules are derived from the complete data set with no validation on samples. Previously, Association rule based Modified Particle Swarm Optimization and Least Squares Support Vector Machine classification is introduced with the capability to lessen the number of rules, looks for association rules on a training set and at last validates them on an independent test set. On the other hand, it only employs categorical data. In case of Type-II Diabetes Mellitus medical diagnosis, the exploitation of continuous data might be essential. With the aim of solving this complication, Improved Frequent Pattern Growth (IFP-Growth) with Hybrid Enhanced Artificial Bee Colony-Advanced Kernel Support Vector Machine (HEABC-AKSVM-IFP Growth) classification based Association Rule Mining (ARM) system is proposed in this study to create rules. This study introduces improved FP-growth to effectively derive frequent patterns including from a vague database in which items possibly will come into view in medical database. Then, HEABC-AKSVM-IFP Growth classifier is employed to create the association rules from the frequent item sets, also keeping away from the rule redundancy and inconsistencies at the time of mining process. Then, results are simulated and evaluated against few classification techniques in terms of classification accuracy, number of derived rules and processing time.

Keywords: Advanced kernel support vector machine, association rule mining, enhanced artificial bee colony, frequent patterns, improved frequent pattern growth algorithm, type-II diabetes

INTRODUCTION

Diabetes mellitus is a major problem and continues to affect many people and its prevention and efficient treatment are unquestionably fundamental. The most common form of diabetes is diabetes type II or Diabetes Mellitus Type (T2DM). In case of diabetes type II, the body is resistant to the consequences of insulin. Few millions of people have been identified with diabetes type II and regrettably, several people are ignorant that they are at high danger. Regardless of current medical advancements, premature diagnosis of disease has developed however almost half of the patients of type II diabetes are ignorant of their disease and possibly will take more than ten years as the delay from disease beginning to diagnosis at the same time early diagnosis and treatment of this disease is very important. Classification systems have been extensively employed in medical domain to discover patient's data and derive a predictive model. They assist physicians to advance their prognosis, diagnosis or treatment planning methods.

In recent past, several investigations have been carried out in the diagnosis of diabetic disease. Numerous techniques with knowledge discovery approaches (Kumari and Singh, 2013), for instance, logistic regression, Naive Bayes, Semi-Naive Bayes, Multi-Layer Perceptrons (MLPs), General Regression Neural Networks (GRNNs), Support Vector Machines (SVMs), Radial Basis Functions (RBFs), Least Square Support Vector Machines (LS-SVMs), Artificial Neural Network (Oner *et al.*, 2013) and Visualization (Guo *et al.*, 2012) have been exploited in some of these investigations. Decision tree schemes also have been extensively exploited to construct classification models as such models strongly look like human reasoning and are simple to understand. Association Rule Mining (ARM) portrays how two items are correlated by means of a special method of exploring patterns different from other analysis approaches. The association rule generated from ARM can originate the relation among X and Y in the form of "$X \rightarrow Y$" or "If X.., then Y..," and analyze it as "If item X exists, item Y coexists".

A rule does not automatically mean cause and effect. Rather, it recognizes simultaneous occurrence

---

Corresponding Author: K. Vembandasamy, PSG College of Arts and Science, Coimbatore, India between items in antecedent X and consequent Y. ARM makes it feasible to examine the association between not only two diseases, however also among three or more comorbidities that can be computed from existing statistics. As a result, the current investigation was conducted to find out the associations among complications, the several diseases that come with T2DM and three or more comorbidities, by means of ARM based on huge amounts of clinical data. Through the skin temperature measured in several components of the body specifically nose, tibia, toe and with Non-contact digital thermography, statistical examination and bio-chemical blood analysis, a survey was carried out for screening the patients with diabetics (Anburajan *et al*., 2011). Fuzzy disease ontology is exploited to model the diabetes knowledge (Lee and Wang, 2011). Current investigation found that 80% of problems can be prevented by identification intelligent data analysis approach like machine learning technique are helpful in identification which can increase in premature detection (Barakat *et al*., 2010).

Karthikeyan *et al*. (2013) employed the hierarchical K-Means clustering approach that is exceedingly utilized with protein dataset and generated successful results. However, when enhancing the amount of inner cluster distance, there is considerable decrease in sum of clusters. A hybrid binary classification model with the fundamental concepts of soft computing and ANN was formulated in Bansal *et al*. (2011). The prediction of being suffering by diabetes with the help of data from EMR is something innovative which demonstrates how transformations in lifestyle like eating habits and doing exercises is significant in the growth of T2DM. This investigation points out that the prediction of being suffering by T2DM from EMR is efficient for developing supportive models (Sankaranarayanan and Perumal, 2014). Decision tree dependent prediction model was formulated in Al Jarullah (2011). In Karthikeyan and Vembandasamy (2014) with the help of enhanced FP-growth the frequent item sets are produced where the Ant Colony Optimization approach is exploited to optimize the frequent item sets produced by Enhanced FP-Growth algorithm. Although this is efficient there is a requirement for alternative effective classifier to guide the rule mining process that could circumvent the missing important rules produced by ARM approaches. In the most of earlier investigations, uncertainty against the constraints of patient's representation has not been considered. In addition, in several existing approaches, the derived results were not appropriate concerning the time or the diagnosis and consequently operate slowly.

Having reviewed the existing literature, the researchers found that the ARM techniques are more accurate than the other methods and lead to better results. In this study, Improved FP-Growth algorithm is employed for generating frequent itemsets. By using this result, the classification is done based on the proposed Hybrid Enhanced Artificial Bee Colony-Advanced Kernel Support Vector Machine to improve the accuracy of rule generation.

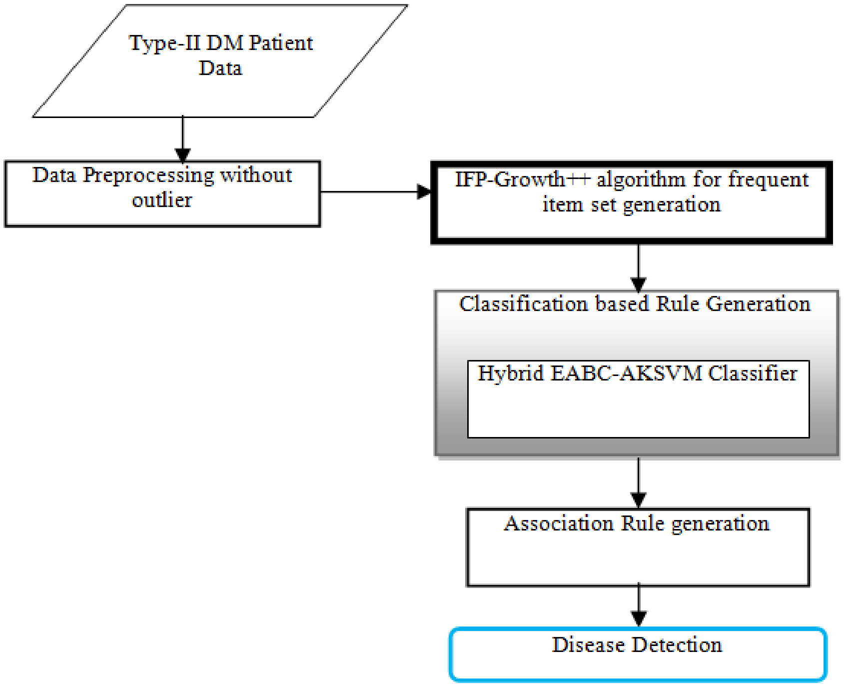

Fig. 1: Block diagram for an intelligent type-II diabetes mellitus diagnosis system

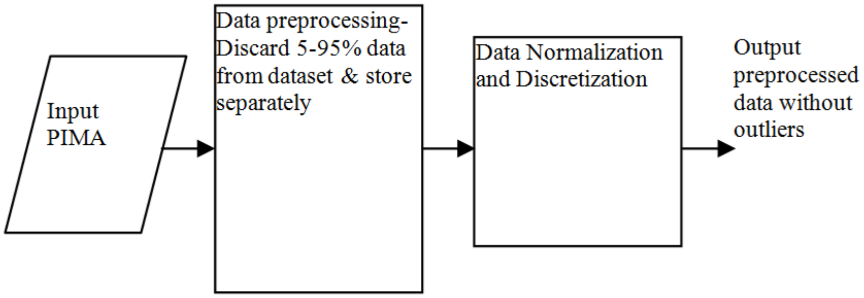

Fig. 2: Data preprocessing with outlier detection and removal

MATERIALS AND METHODS

The Type 2 Diabetes (TTD) is a most common kind of diabetes and accounts for 90-95% of all diabetes. Detection of TTD from classified factors or symptoms has grown to be a matter of concern, which is not free from false presumptions because of unpredictable effects. In conjunction with this situation, the data mining could be acclimated, recommends us in competent investigation from data. This study proposes Hybrid Enhanced Artificial Bee Colony with Advanced Kernel Support Vector Machine (HEABC-AKSVM-IFP Growth) in the data mining accomplishment to access information from authentic data of patient medical records. It provides a decision-making assistance through association rule mining based classification with the frequent itemset result of Improved FP-Growth.

Figure 1 demonstrates the block diagram of the proposed system. At the beginning, the Type-II DM patient data is provided as input. This possibly will contain unnecessary and empty data, this considerably diminish the detection accuracy. In subsequent step, preprocessing is carried out to eradicate the noisy data. This is completed through the Normalization without outliers and Chi Merge discretization approach. Following that, IFP-growth algorithm is employed for discovering frequent patterns of the input dataset. The resultant frequent itemsets from IFP-Growth is then given to HEABC-AKSVM classification algorithm, thus, the accurate result of association rules are generated. The detailed explanation is provided in the following sections.

Dataset pre-processing: The PIMA Indian diabetes dataset was contributed by Vincent Sigillito and it includes 768 instances. Especially, the entire patients here are females with at least 21 years old of Pima Indian heritage. Pre-processing is an extremely important phase, since it will influence the result of a classification algorithm. This module computes tuples with missing values by means of several options like maximum, minimum, constant, average and standard deviation for the handling of missing values tuples prior to executing the normalization approach on the dataset.

Table 1: Mean and standard deviation before normalization

Attributes No.	Mean	S.D.
a_1	3.8	3.4
a_2	120.9	32.0
a_3	69.1	19.4

S.D.: Standard deviation

This step works in two phases. During the initial phase, after pre-processing the dataset, calculate and remove 5-95% data from the dataset. Accumulate and normalize these removed data separately which is regarded as outliers. The outliers' detection is a process that discovers objects that are unrelated or inconsistent with respect to the other data. It has several advantages in applications like fraud detection, network intrusion detection and clinical diagnosis of diseases. As a result, the analysis confirms that outlier detection and removal will produce the best and most effective and precise results using preprocessing (Patel and Mehta, 2011). This process is illustrated in Fig. 2.

Normalization: Normalization is the process of scaling of data transformation of instances. In a dataset, the instance possibly will have several values ranging from smallest to highest values. This step normalizes the entire numerical values exist in the dataset. The result values engage the range between (0, 1). With the help of translation and scale parameters, the range can be modified. Equation (1) provides the computation of Z-Score normalization technique. In Eq. (1), $\bar{A}_n$ and σA_n v represents the mean, standard deviation and value of attribute $A = \{a_1, a_2, \ldots a_n\}$. In this study, the attribute A value is given as 13:

$$v' = \frac{\sum_{i=1}^{n} v-\bar{A}_n}{\sum_{i=1}^{n} \sigma A_n} \tag{1}$$

Mean and standard deviation of the attributes in the PIMA dataset prior to normalization are provided in Table 1. At this point, for instance, consider the initial three attributes namely a_1, a_2 and a_3. Each have distinct mean values and standard deviation, in order that, a_1 has lowest mean value (3.84), a_2 has higher value (120.8) and a_3 has medium value (69.1) and the difference among these mean values of each of the attributes is diverse highly as given in Table 1.

Table 2: Mean and standard deviation after normalization

Attributes No.	Mean	S.D.
a_1	0.226	0.19
a_2	0.608	0.16
a_3	0.566	0.15

S.D.: Standard deviation

Table 3: Results of a_1 after discretization

Label	Count	TP	TN
(-inf-0.382)	599	173	426
(0.382-inf)	166	95	74

At this time, the attribute values normalized by means of the given formula: $x_{mean} => (x - Mean)/Deviation$ and $x_{std} - deviation => (x - Deviation)/Mean$. Following the normalization process, the mean and standard deviation values are normalized to be present in the range -1 to $+1$ and the difference among each attribute is less as shown in Table 2.

Numerical data discretization: Discretization is the method of transforming the quantitative data into qualitative data. Quantitative data are more frequently provided in several data mining applications. But learning algorithms typically takes qualitative data into account; with a quantitative data, the learning algorithm performs less competent and less effective values. The swiftness of proposed algorithm can be enhanced with the help of discretized variables. Attributes after discretization all attributes are given for a_1 in Table 3. Count indicates the label count; TP (tested_positive) provides the number of data with tested_positive values in the count. TN (tested_negative) indicates the number of data with tested_negative values in the count. In this study, Chi Merge is employed as a discretization method given in "(Karthikeyan and Vembandasamy, 2015)".

In this study, HEABC-AKSVM-IFP Growth has been formulated, that is a new dimension into the data mining approach. To facilitate this, the conception of classification based association rules have been incorporated into the improved Frequent Pattern (FP) method. This proposed approach works better than conventional association rule mining approaches and IFP skilfully performs in large uncertain database also. The proposed algorithm has two steps: First is frequent itemset generation and, second is classification based rule generation.

Frequent itemset generation: The proposed IFP-Tree includes primarily two elements-the tree and a table. The tree indicates the correlation between the items more distinctively and table is exploited to accumulate the spare items. It is regarded as spare table (Stable) which has two columns as item_name and frequency. Item_name indicates the name of the items and frequency indicates how many times it takes place in Stable. The major reason to commence the spare table is, in conventional FPtree numerous branches are generated and the identical item comes out in more than one node. However in the proposed IFP-tree, each distinctive item has only one node. Consequently, it is simpler and well-organized for further processing. The complete process of frequent itemset generation is provided in Algorithm 1.

Lemma 1: Consider A is the set of items. Spare item S is defines as $S \subset A$. There are two scenarios where an item is regarded as spare.

Case 1: If the item contains no single edge from the current root to its node. That indicates the item is previously exists in the IFP-tree. In proposed IFP-tree, the root is not kept constant, it changes.

Case 2: The transaction items that do not include the most common item.

Lemma 2: The frequency of an item in the IFP-tree is the amount of time it takes place in the IFP-tree. It is given as (item_name: frequency of IFP-tree).

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

Lemma 4: An uncertain transaction t is a transaction that includes uncertain items. A transaction database containing uncertain transactions is known as an uncertain transaction Database (D).

Case 1: In the process of building the UIFP tree (Uncertain Improved FP), the frequent 1-itemsets with their expected support ($expSup$) higher than or equal to the minimum count were initially found. Each transaction in the uncertain database was then revised to maintain only the frequent 1-items (if $expSup(I) = n \times s$). The construction process of the UIF tree was a slightly like constructing an IFP-tree apart from that the same items with diverse existential probabilities was inserted in different nodes. Only the same items with the identical existential probabilities in transactions were combined collectively in the tree.

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, $S, C, F, R, j, n, K_1, expSup$)
```

```
for each item $i$ in FP-tree where $(i! = R)$ do
if $i.S = i.F$ then
frequency of the frequent item set, $K = i.F$
Create item set, $P = (i\mu)$ with the frequency value of the tree.
/* μ = Every possible combinations of the item and nodes with higher frequency in FP-tree */
Frequent item set is written in $\{P: K\}$ format.
else if $i.S > i.F$ then
    Frequency of frequent item set $K = i.F + C$
    \* $C$ =frequency in Stable count*/
Generate item set, $P = (ia)$
/*$a$ = Every possible combination of item and each intermediate nodes up to most frequent item node in IFP-tree*/
else Frequency of the frequent item set $K = i.F$
Create item set, $P = (i\beta)$
for each number of transactions $n \geq ++j$ do
for each frequent item $i \in K_1$ do
if $expSup(i) = n \times s$
then $i$ is an uncertain frequent-1 item sets
end for
end for
```

At this moment, it is important that in every previous approach to discover the frequent item set generated like (Tudor, 2008) 1-itemset, 2-itemset and so on. On the other hand, in this approach, all categories of possible item set which meet the minimum user defined support is generated. This enables the proposed approach efficient than the other approaches. At this instant, it is relatively simple to define the association rules from the frequent item set. By means of generating additional array (expAry) for each item in uncertain transactional database to maintain the information of its super-item sets in CUIFP-mine, redundant node generation and the computational cost in the mining process is significantly reduced.

The rule generation stage: The frequent itemsets generated are exploited to produce association rules that satisfy the minimum support and minimum confidence. Normally, in association rule mining, whichever item that meets minsupp is known as a frequent itemset. The frequent itemsets are produced with the help of IFP-Growth explained in above section. Once the frequent items have been found out, classification depending on association rules algorithms obtain a complete set of Class Association Rules (CAR) for those frequent items that satisfy minconf. The major task of CAR-RG (Class Association Rule-Rule Generation) is to discover all ruleitems that have support higher than minsup. A ruleitem is of the form: $< I, y >$ where I represents a set of items, $y \in Y$ represents a class label. The support count of the I (called $Isupcount$) is the number of cases in D that contain the I. The support count of the ruleitem (called $rulesupCount$) represents the number of cases in D that include the I and are labeled with class y. Each ruleitem principally represents a rule: $I \rightarrow y$, whose support is $(rulesupCount\ / |D|\ ) * 100\%$, in which $|D|$ represents the size of the dataset and whose confidence is $(rulesupCount/ IsupCount) * 100\%$. Ruleitems that meet minsup are called frequent ruleitems, at the same time the rest are called infrequent ruleitems.

For instance, consider the following as a ruleitem:

$$< \{(A,1),(B,1)\},(class,1) >$$

where, A and B represents 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 meets the minsup criterion, can confirm it is frequent. For the entire ruleitems that have the similar 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 similar maximum confidence, then arbitrarily choose one ruleitem. For instance, consider two ruleitems that have the similar I: 1. $< \{(A,1),(B,1)\}, (class{:}\,1) >$ 2. $< \{(A,1), (B,1)\}, (class{:}\,2) >$. Consider the support count of I 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 higher than minconf, can confirm the rule is accurate. The set of Class Association Rules (CARs) accordingly includes the entire PRs that are both frequent and accurate. In order to accomplish this process, a HEABC-AKSVM-IFP Growth classifier with high accurate results is proposed.

Advanced kernel support vector machine: Let M m-dimensional training inputs $x_i\ (i = 1,\ldots,M)$ is a member of Class 1 or 2 and the related labels be $y_i = 1$ for Class 1 and -1 for Class 2. When these data are linearly separable, the decision function is given as:

$$y_i\ (w^T x_i + b\ ) \geq 1\ for\ i = 1,\ldots,M$$

The decision boundary can be determined by solving the following constrained optimization problem:

$$\begin{aligned}&Minimize\ \tfrac{1}{2}\|W\|^2\\&Subject\ to\ y_i\ (w^T x_i + b\ ) \geq 1\end{aligned}$$

In case of non linearly seperable setback, introduce a slack variable ξ. When error is between $0 \leq \xi \leq 1$, data can be accurately classified, however when $\xi \geq 1$, at that time the data is misclassified. Hence, ξ should be minimized. The hyperplanes are calculated as:

$$y_i\ (w^T x_i + b\ ) \geq 1 - \xi_i, for\ y_i = 1$$

$$y_i\ (w^T x_i + b\ ) \leq -1 + \xi_i, for\ y_i = -1$$

The decision boundary can be determined by the following optimization setback:

Minimize $\frac{1}{2}\|W\|^2 + C\sum_{i=1}^{M}\xi_i$

Subject to $y_i\ (w^T x_i + b\ ) \geq 1 - \xi_i$ for $\xi_i > 0$

Training vectors x_i are mapped into a higher (possibly be infinite) dimensional space by the function *Φ*. At that moment, SVM discovers a linear separating hyperplane with the maximal margin in this higher dimension space $C > 0$ is the penality parameter of the error term. In addition, $K\left(x_i, x_j\right) \equiv \Phi\left(x_i\right) T\ \Phi\left(x_j\right)$ is known as the kernel function. In case of statistical classification, the Fisher kernel, named in credit of Sir Ronald Fisher, is a function that determines the similarity of two objects on the basis of sets of measurements for every object and a statistical model. In a classification process, the class for a new object (whose real class is indefinite) can be estimated by diminishing, across classes, an average of the Fisher kernel distance from the new object to every recognized member of the particular class. The Fisher kernel was formulated in 1998 and it integrates the merits of generative statistical models (like the hidden Markov model) and those of discriminative approaches (like support vector machines): generative models can process data of inconsistent length (adding or removing data is well-supported) and discriminative approaches can have flexible criteria and provide better outcome. The Fisher kernel makes use of the Fisher Score, given as $U_X = \nabla_\theta\ log\ P\ (X|\theta)$ with θ being a set (vector) of parameters. The function taking θ to $log\ P\ (X|\theta)$ represents the log-likelihood of the probabilistic model. The Fisher kernel is given in Eq. (2) with I indicates the Fisher information matrix as:

$$K\left(X_i, X_j\right) = U_{x_i}^T I^{-1} U_{x_j} \tag{2}$$

Enhanced artificial bee colony: With the intention of obtaining optimized hyper parameters of AKSVM, the newly developed Swarm Intelligent technique, ABC is incorporated into AKSVM. ABC approach was proposed by Karaboga for real parameter optimization. In ABC approach, when the solution space of the problem is D-dimensional, where D represents the number of parameters to be optimized. In this study, the parameters involved are w^T, b and θ. The fitness value of the arbitrarily selected site is provided in Eq. (3) as given below:

$$fit_i = \frac{1}{(1 + obj.Fun_i)} \tag{3}$$

The size of Employed Bees (EB) and Onlooker Bees (OB) are equally SN (swarm of food sources (SN)), which is equivalent to the amount of food sources. For every food source's location, one employed bee is allocated to it. For every employed bee whose overall numbers are equivalent to the amount of the food sources, a new source is obtained based on Eq. (9):

$$v_{ij} = x_{ij} + \varphi_{ij}\left(x_{ij} - x_{kj}\right) \tag{4}$$

where, $i = \{1, 2, \ldots, SN\}$, $j = \{1, 2, \ldots, D\}$, φ represents a random generalized real number inside the range $(-1,\ 1)$. k represents a randomly chosen index number in the Bee colony. Following to the production of new solution $v' = \left\{x_{i1}^{'}, x_{i2}^{'}, \ldots, x_{iD}^{'}\right\}$, it is evaluated to the original solution $v = \{x_{i1}, x_{i2}, \ldots, x_{iD}\}$. When the new solution is better than previous one, the bee remembers the new solution; if not bee remembers the former solution. The onlooker bee chooses a food source to utilize with the probability provided in Eq. (5):

$$P_i = \frac{fit_i}{\sum_{j=1}^{SN} fit_j} \tag{5}$$

where, fit_i indicates the fitness of the solution v. SN represents the number of food sources positions. Afterwards, the onlooker bee looks a new solution in the chosen food source site by Eq. (9), the similar way as exploited by employed bees. In Scout Bee (SB) phase, when the fitness of a discovered food source hasn't been enhanced for a specified number of trial (indicated by limit), it is discarded. This task indicates the negative feedback in ABC algorithm and the EB of that food source turns into a SB and makes an arbitrary search by Eq. (6):

$$x_{id} = x_d^{min} + r\left(x_d^{max} - x_d^{min}\right) \tag{6}$$

where, $r = a$ is a random real number inside the range $(0, 1)$. x_d^{min} and x_d^{max} represents the lower and upper borders in the d^{th} dimension of the problems space.

Hybrid EABC-AKSVM classifier: This section discusses the CAR-CB (Classification Association Rule-Classifier Bulider) approach for constructing a classifier with the help of CARs. In order to construct the finest classifier out of the complete set of rules would involve evaluating all the feasible subsets of it on the training data and choosing the subset with the right rule sequence that provides the smallest amount of errors. There are 2^m such subsets, in which m represents the number of rules, which can be more than 10,000, not to point out different rule sequences. This is obviously infeasible. The classifier it constructs performs extremely well as compared to that constructed by MPSO-LS-SVM. Since a new member

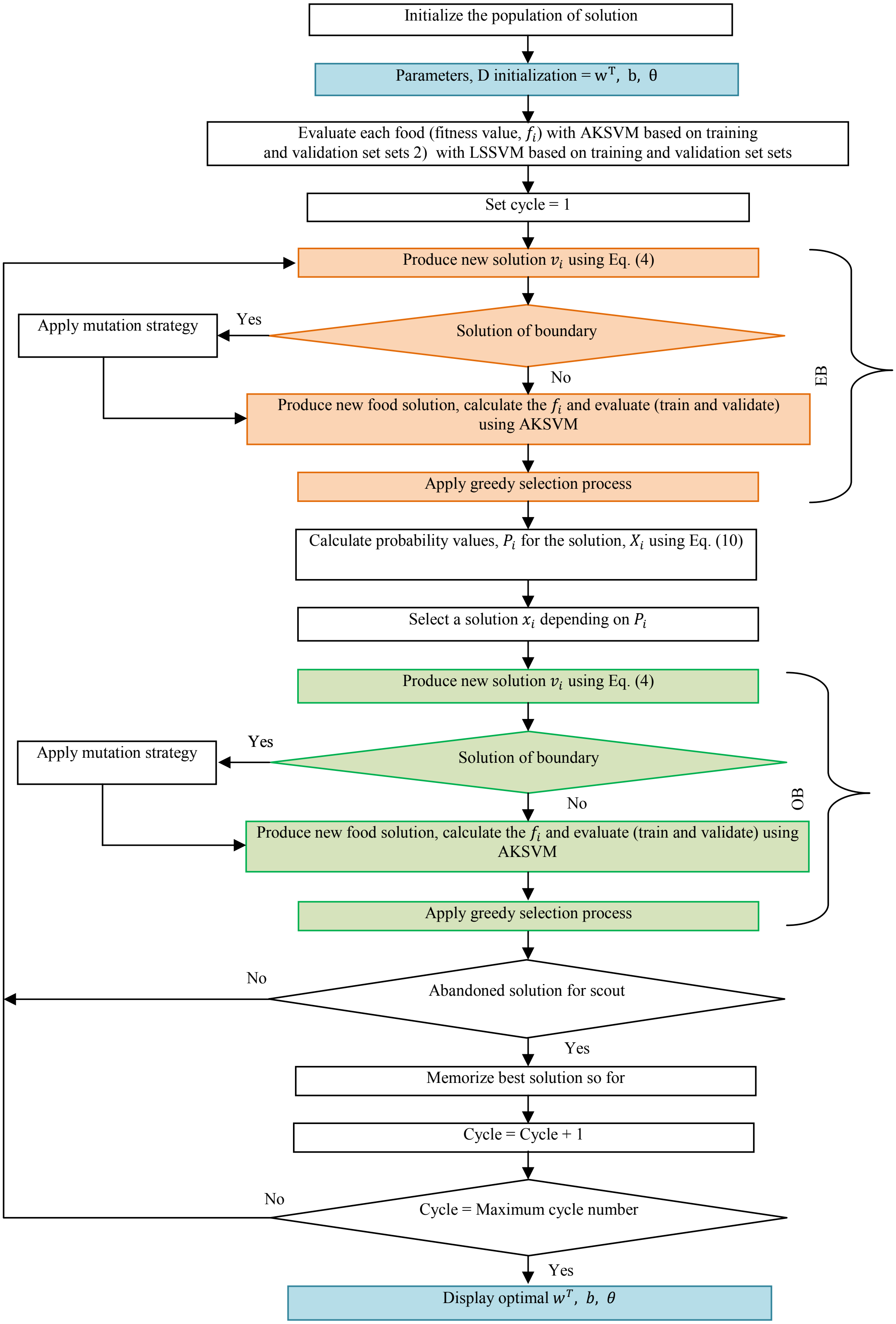

Fig. 3: Flow diagram of HEABC-AKSVM

of Meta heuristic family, ABC has confirmed to be an effective approach in solving optimization problem. As compared to GA and Ant Colony Optimization (ACO), ABC revealed its potential in producing outstanding performance (Shokouhifar and Abkenar, 2011).

On the other hand, occasionally, ABC operates too well. When safeguard step is not taken into account, ABC model inclined to converge too quick and this possibly will lead to local minima. This might occur when the area explored by the model is not a preferred area, specifically, global minima area. As a result, an improvement in ABC is made by implementing mutation approach (Haupt and Haupt, 1998). By including mutation scheme in ABC, the model is provoked to discover other areas with the aim of searching for global minima rather than local minima. When the unnecessary circumstances aforementioned occurred, rather than forcing the parameter value to the boundary, mutation strategy is commenced. This task is implemented multiplying the generated random number with the range of boundary that has been determined. In this study, the boundaries are fixed to the range of between (1, 1000):

$$new\_param = (ub - lb) * rand\_num \qquad (7)$$

where, new_{param} = new parameter, $rand\_num$ = random number, ub = upper bound, lb = lower bound. Figure 3 illustrates the uncomplicated form of the proposed prediction model at the same time the flow of Hybrid EABC algorithm with AKSVM. From this figure, it can be revealed that the mutation process is implemented in both employed and onlooker bees phases before constructing new food solution.

At last, the refined rule produced by IFP-Growth-Hybrid EABC-AKSVM is given below for example, which is exploited for the prediction and source 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%.

RESULTS AND DISCUSSION

The Pima Indian Diabetes Dataset is freely available at UC Irvine Machine Learning Lab and extensively employed as a standard for testing the accuracy of diabetes status by means of data mining algorithms (http://www.uci.edu/). The dataset includes 512 training examples and 256 examples as test data with 8 attributes. The attribute are provided in Table 4. The 13th attribute is a diabetic class which includes two values tested positive and tested negative of nominal type. Out of 768 patients 500 are found negative (Class = 0) and 268 are found positive (Class = 1).

Following the preprocessing phase discussed in above section, executed CAR-RG based on HEABC-AKSVM-IFP Growth to produce the refined rules. The results of proposed HEABC-AKSVM-IFP Growth are compared against the existing ABC-LSSVM-IFP Growth, MPSO-LSSVM-CFP Growth++ and SVM-FP Growth based classification technique based on accuracy rate, runtime and number of rules produced etc.

Convergence performance: The comparison in terms of convergence between HEABC-AKSVM-IFP Growth, ABC-LSSVM-IFP Growth, MPSO-LSSVM-CFP Growth++ and SVM-FP Growth is illustrated in Fig. 4. The efficiency of proposed model can be seen from the narrow span showed in the graph, where the predicted value by HEABC-AKSVM-IFP Growth is more accurate than ABC-LSSVM-IFP Growth, MPSO-LSSVM-CFP Growth++ and SVM-FP Growth. In addition, the proposed model also offers better convergence performance as compared to standard ABC-LSSVM.

Average number of identified rules comparison: The comparison in terms of average number of rule generation between HEABC-AKSVM-IFP Growth, ABC-LSSVM-IFP Growth, MPSO-LSSVM-CFP

Table 4: 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) (weight in kg/ (height in m)^2)	Numeric
a_7	PEDI diabetes pedigree function	Numeric
a_8	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 diagnose results (“tested_positive”, -tested_negative”)	Nominal

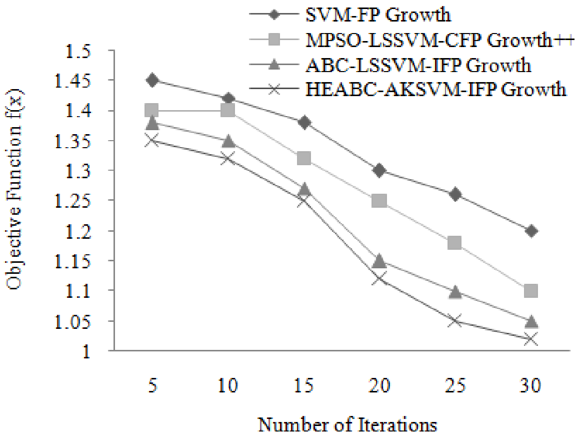

Fig. 4: Convergence performance

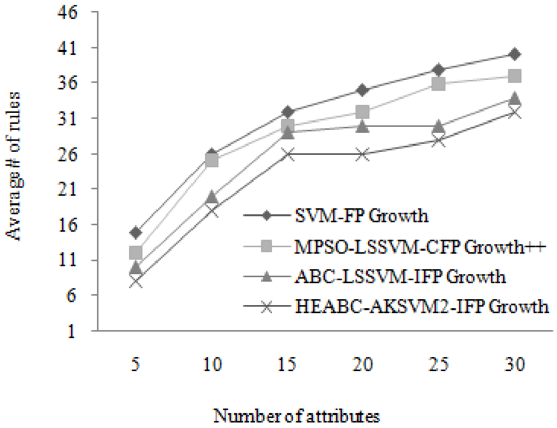

Fig. 5: Average number of identified rules comparison results

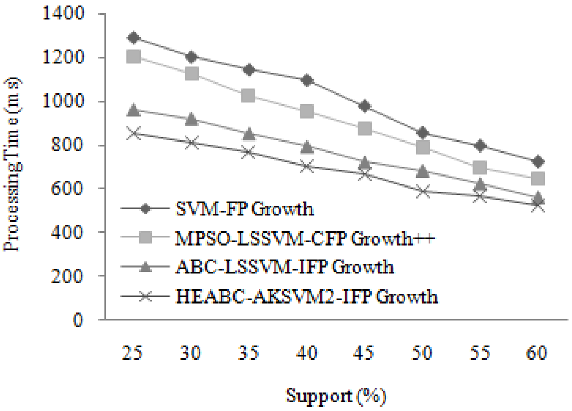

Fig. 6: Processing time comparison

Growth++ and SVM-FP Growth is illustrated in Fig. 5. When the number of attributes is increases the average number of rules is increases. However, average number of rules generated by HEABC-AKSVM-IFP Growth 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 by HEABC-AKSVM-IFP Growth is more accurate than ABC-LSSVM-IFP Growth, MPSO-LSSVM-CFP Growth++ and SVM-FP Growth.

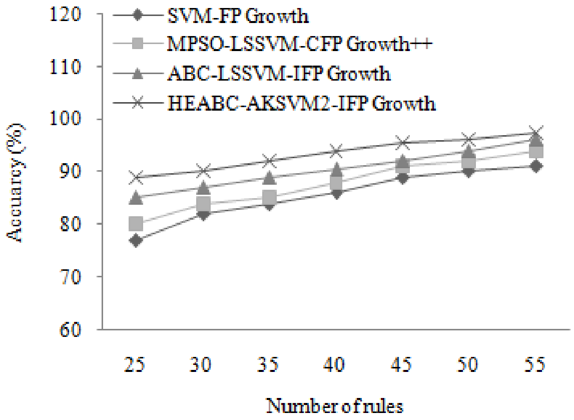

Fig. 7: Accuracy comparison

Processing time comparison: The IFP-Growth algorithm discovers frequent item sets and Fig. 6 shows much greater efficiency than existing algorithm ABC-LSSVM-IFP Growth, MPSO-LSSVM-CFP Growth++ and SVM-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 HEABC-AKSVM is reportedly working efficiently and in many cases, it's much faster than ABC-LSSVM-IFP Growth, MPSO-LSSVM-CFP Growth++ and SVM-FP Growth.

Accuracy comparison: The IFP-Growth algorithm discovers frequent item sets and Fig. 7 shows much greater accuracy results than existing algorithm ABC-LSSVM-IFP Growth, MPSO-LSSVM-CFP Growth++ and SVM-FP Growth. When the number of rules generation is increases, the accuracy of the result is also increases. The proposed algorithm produces high accuracy rate when compared to existing system. The algorithm IFP-Growth with HEABC-AKSVM is working effectively and in many cases produces high accuracy rate than ABC-LSSVM-IFP Growth, MPSO-LSSVM-CFP Growth++ and SVM-FP Growth.

CONCLUSION

An intelligent HEABC-AKSVM-IFP Growth technique is proposed in this study for the better prediction of type 2 diabetes mellitus disease. This study gathers and investigates medical patient record of type 2 diabetes mellitus with data mining approaches to obtain the information from TTD patient. Initially the preprocessing phase is done, the frequent itemsets are generated using IFP-Growth, then executed CAR-RG based on HEABC-AKSVM to produce the refined rules. The storage capacity and computation cost is reduced through this method. This approach is extremely faster and more efficient in terms of predicting the rules and producing the rules. The experiment has been successfully carried out using Pima Indian Diabetes Dataset is freely available at UC Irvine Machine Learning Lab and evaluated the proposed HEABC-AKSVM-IFP Growth against the existing ABC-LSSVM-IFP Growth, MPSO-LSSVM-CFP Growth++ and SVM-FP Growth based classification technique based on accuracy rate, runtime and number of rules produced etc. HEABC-AKSVM-IFP Growth as part of data mining technique achieves better performance and effectively handles very large size of data set than other existing methods such as ABC-LSSVM-IFP Growth, MPSO-LSSVM-CFP Growth++ and SVM-FP Growth. Rules are extracted using HEABC-AKSVM-IFP Growth based Association Rule Mining can be used by physician to diagnose TTD disease. In future, an effective approach can be formulated to derive association rule from the database with missing values without generating candidate itemsets.

REFERENCES

Al Jarullah, A.A., 2011. Decision tree discovery for the diagnosis of type II diabetes. Proceeding of IEEE International Conference on Novations in Information Technology (IIT, 2011), pp: 303-307.

Anburajan, M., S. Sivanandam, S. Bidyarasmi, B. Venkatraman, M. Menaka and B. Raj, 2011. Changes of skin temperature of parts of the body and serum asymmetric dimethylarginine (ADMA) in type-2 diabetes mellitus Indian patients. Proceeding of 33rd Annual International Conference of the IEEE EMBS, pp: 6254-6259.

Bansal, J.C., P.K. Singh, M. Saraswat, A. Verma, S.S. Jadon and A. Abraham, 2011. Inertia weight strategies in particle swarm optimization. Proceeding of 3rd World Congress on in Nature and Biologically Inspired Computing (NaBIC), pp: 633-640.

Barakat, N.H., A.P. Bradley and M.N.H. Barakat, 2010. Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE T. Inf. Technol. B., 14(4): 1114-1120.

Guo, Y., G. Bai and Y. Hu, 2012. Using bayes network for prediction of type-2 diabetes. Proceeding of the IEEE International Conference for Internet Technology and Secured Transactions, pp: 471-472.

Haupt, R. and S. Haupt, 1998. Practical Genetic Algorithms. Wiley-Interscience Publication, New York.

Karthikeyan, T. and K. Vembandasamy, 2014. A refined continuous ant colony optimization based FP-growth association rule technique on type 2 diabetes. Int. Rev. Comput. Software, 9(8): 1476-1483.

Karthikeyan, T. and K. Vembandasamy, 2015. A novel algorithm to diagnosis type II diabetes mellitus based on association rule mining using MPSO-LSSVM with outlier detection method. Indian J. Sci. Technol., April, 8(S8): 310-320.

Karthikeyan, T., R. Ragavan and K. Vembandasamy, 2013. Hierarchical K-means clustering algorithm for an E-care of diabetes mellitus. Int. J. Adv. Res. Comput. Sci. Software Eng., 3(12): 653-660.

Kumari, S. and A. Singh, 2013. A data mining approach for the diagnosis of diabetes mellitus. Proceeding of 7th International Conference on In Intelligent Systems and Control (ISCO, 2012), pp: 373-375.

Lee, C.S. and M.H. Wang, 2011. A fuzzy expert system for diabetes decision support application. IEEE T. Syst. Man Cy B, 41(1): 139-153.

Oner, Y., T. Tunc, E. Egrioglu and Y. Atasoy, 2013. Comparisons of logistic regression and artificial neural networks in lung cancer data. Am. J. Intell. Syst., 3(2): 71-74.

Patel, V.R. and R.G. Mehta, 2011. Impact of outlier removal and normalization approach in modified k-means clustering algorithm. Int. J. Comput. Sci. Issues (IJCSI), 8(5): 331-336.

Sankaranarayanan, S. and T.P. Perumal, 2014. A predictive approach for diabetes mellitus disease through data mining technologies. Proceeding of 2014 World Congress on Computing and Communication Technologies (WCCCT), pp: 231-233.

Shokouhifar, M. and G.S. Abkenar, 2011. An artificial bee colony optimization for MRI fuzzy segmentation of brain tissue. Proceeding of the International Conference on Management and Artificial Intelligence. Indonesia, pp: 6-10.

Tudor, I., 2008. Association rule mining as a data mining technique. B. U. Petrol-Gaze Din Ploiesti, LX(1): 49-56.