ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 8(S8), 310–320, April 2015* DOI: 10.17485/ijst/2015/v8iS8/53631

A Novel Algorithm to Diagnosis Type II Diabetes Mellitus Based on Association Rule Mining Using MPSO-LSSVM with Outlier Detection Method

T. Karthikeyan* and K. Vembandasamy

P.S.G. College of Arts and Science, Coimbatore, India; t.karthikeyan.gasc@gmail.com, vembupsgphd@gmail.com

Abstract

Background/Objectives: The execution of Frequent Pattern Growth algorithm on medical data is difficult. Association rule based classification is an interesting area focused that can be utilized for early diagnosis. **Methods/Statistical analysis**: Discretization phase is necessary to transform numerical characteristics. The results are given to Complete Frequent Patten Growth++ for the purpose of rule induction. Accordingly, using Modified Particle Swarm Optimization together with Least Squares Support Vector Machine scheme (MPSO-LSSVM) rules are produced with outlier detection method. Pima Indians Diabetes Data Set is taken as an input. The execution time, number of rules generation and the detection of outlier percentage are analyzed. **Results:** The CFP-growth algorithm utilizes for finding frequent patterns where constructing the Minimum Item Support (MIS)-tree, CFP-array and producing frequent patterns from the MIS-tree. From the set of frequent item sets found, create all the association rules that have a confidence exceeding the minimum confidence. The Enhanced outlier detection method is used for determining the outlier degree from association rules for outlier detection. Association rules are mined using MPSO-LSSVM classification based association rule mining algorithm. The classification based association rule generation using MPSO-LSSVM is utilized first time in this work with outlier detection method. For the reason of eradicating the effect of unavoidable outliers in investigation sample on a scheme's performance, a new MPSO-LSSVM with the integration of outlier detection method is proposed first time. The experimental observations reveal that this framework provides a better accuracy of 95% when evaluated against the existing techniques. **Conclusion/ Application:** CFP-Growth++ proposed for rule pruning and MPSO-LSSVM based algorithm used for mining association rules from Type-2 DM dataset. This work is suitable for early detection of type-2 diabetes mellitus disease.

Keywords: Association Rule Discovery, Classification, Complete Frequent Patten Growth, Feature Selection, Outlier Detection Approach

1. Introduction

Diabetes Mellitus (DM) is a collection of metabolic infections in which a human being has elevated blood sugar, either for the reason that the pancreas does not generate sufficient insulin, or because cells don't react to the insulin that is generated. This elevated blood sugar makes the conventional signs of polyuria (regular urination), polydipsia (increased need for liquids) and polyphagia (increased starvation). DM includes three major categories, "Type I DM", as a consequence of the human body's malfunction to generate insulin, and necessitates the individual to insert insulin or carry an insulin pump. This category was previously indicated as "Insulin-Dependent Diabetes Mellitus" (IDDM). The second category of DM is recognized as "Type II DM" as a consequence of insulin confrontation, a situation in which cells are ineffective to exploit insulin appropriately, occasionally merged with an absolute insulin insufficiency. This category also called as "Non Insulin

**Author for correspondence*

Dependent Diabetes Mellitus" (NIDDM) or "adult-onset diabetes". At last, "gestational diabetes" takes place when conceived women without an earlier diagnosis of diabetes increase high blood glucose intensity; it possibly will lead to development of type I DM.

Additional categories of DM comprises of congenital diabetes, which is because of genetic deficiencies of insulin discharge, cystic fibrosis-associated diabetes, steroid diabetes stimulated by elevated amount of glucocorticoids, and numerous categories of monogenic diabetes[1]. All categories of DM have few similar characters. In general, the physical body absorbs the sugars and carbohydrates that are consumed and breaks down into a special sugar called glucose. Glucose energizes the cells inside the body. On the other hand, the cells also require insulin, a kind of hormone, in person bloodstream, for the purpose of taking the glucose and utilize it for energy. All categories of DM have been treatable, in view of the fact that insulin became obtainable in 1921. Both type I & II are chronic states that are not possible to treat. Pancreas transplantations have been attempted with some degree of achievement in type I DM, gastric bypass surgical treatment has been doing well in several morbid obesity and type II DM. Gestational DM typically resolves subsequent to delivery[2]. Untreated DM possible will pave way for serious complications. Sensitive impediments consist of diabetic ketoacidosis and non ketotic hyperosmolarcoma. Series long period impediments comprise cardiocascular disease, chronic renal breakdown, and diabetic retinopathy.

Satisfactory dealing of the disease is extremely essential, in addition to blood pressure maintenance and standard of living factors, for example, preventing smoking and keeping a healthy body weight. In view of the fact that the cells can't acquire the glucose, it increases in human blood. Elevated intensities of blood glucose can injure the minute blood vessels in physical kidneys, heart, eyes or nervous scheme, that's the reason diabetes can ultimately basis for heart illness, stroke, kidney disease, sightlessness and harm to nerves in the feet (especially if left untreated)[3]. Association discovery is one of the most common data mining techniques that are used to extract interesting knowledge from large datasets[4]. Much effort has been made to use its advantages for classification under the name of associative classification[5]. Association discovery aims to find interesting relationships between the different items in a database[6,7], while classification intends to realize a model from training data that can be exploited to recognize the class of test patterns. Both association discovery and classification rules mining are essential in practical data mining applications[9] and their integration could result in greater savings and convenience for the user. In [8] the focal pointed largely to discover outlier transactions in transactional database. Two approaches: Association rule and Frequent Pattern (FP) are evaluated for well-organized discovery of outliers. In case of association rule approach, the associative closure model with elevated confidence is characterized and formula for computing the degree of outlier is obtained. In case of FP approach, Apriori algorithm is exploited for the discovery of frequent patterns. On the other hand, discovering such frequent pattern is extremely time consuming process. Rare association rule indicates to an association rule generation among frequent and rare items or among infrequent items. Complete Frequent Patten (CFP) growth is a kind of scheme which is exploited to obtain frequent patterns with the help of multiple minimum support (minsup) values. This scheme is a development of FP-growth technique to multiple minsup values. This scheme engages building up of Multiple Item Support tree (MIS-tree) and producing common patterns from the MIS-tree. The major concern in CFP-growth is building up the compact MIStree, since certain items are taken into account by CFP-growth, which will produce neither common patterns nor rules. In this research, an efficient approach proposed for building the compact MIS-tree. For this purpose, the proposed approach investigates the notions, for instance, least minimum support and infrequent child node pruning. The proposed approach enhances the overall performance over CFP-growth approach.

For heart disease prediction a novel lazy associative classification approach is proposed that is used to assist physicians to get perfect assessments in Andhra Pradesh[16]. Based on interpretable fuzzy association rules and routine membership function generation is achieved through new classification model[17]. To fulfill the efficiency criteria of interpretable fuzzy association rules using a novel classification model[18]. While this model degrades the accuracy when this is affected by the partitioning of numerical attributes, solve the accuracy of the classification model. Pach et al[18] conversed numerous fuzzy and crisp partitioning techniques to solve this accuracy problem. In [19] based on an extended association rule mining technique a new approach is proposed to construct effective classifier. In [20] based on enhanced FP-growth the frequent item sets are generated where the Ant Colony

Optimization algorithm is used to optimize the frequent item sets generated by Enhanced FP-Growth algorithm. Though this is effective there is a need of alternative effective classifier to lead the rule mining process that could avoid the missing important rules generated by association rule mining techniques. In [21] a new direct associative classification method called IGARC proposed that is an improvement of GARC approach, extracts directly generic associative classification rules from a training set in order to reduce the number of associative classification rules without jeopardizing the classification accuracy.

In [22] a new association rule-based text classifier algorithm proposed to improve the prediction accuracy of Association Rule-based Classifier By Categories (ARC-BC) algorithm. In [23] investigated the Electroencephalogram (EEG) of twenty schizophrenic patients and twenty age-matched healthy subjects are analyzed for classification purposes. Several features including AR model coefficients, band power and fractal dimension are extracted from EEG signals. This paper proposes a new classification method based on association rule mining. Fuzzy Accuracy-based Classifier System (F-XCS) is used to improve the resulted fuzzy associative rules for discriminating between healthy and schizophrenic subjects. In [24] introduced ARUBAS, a new framework to build associative classifiers. In contrast with many existing associative classifiers, it uses class association rules to transform the feature space and uses instance-based reasoning to classify new instances. The framework allows the researcher to use any association rule mining algorithm to produce the class association rules. In [25] investigated how the application of a single granularity learning approach influences the performance of fuzzy associative rule-based classifiers. The aim is to reduce the complexity of the obtained models, trying to maintain good classification ability. However, this approach produces models with a slightly decreased accuracy, which is balanced by a considerable reduction of models' complexity.

The major intention of this research is to build a classification scheme for DM analysis and treatment with the help of a hybrid algorithm which includes Modified-PSO algorithm and LS-SVM classifier. Least Squares-Support Vector Machine (LS-SVM) classifier is one of the categories of Support Vector Machine (SVM) model[10]. LSSVM is exploited for making a decision for best possible hyper plane, which splits different classes. It acquires this best possible hyper-plane with the assistance of maximum Euclidean distance to the close point. It is a kind of parametric algorithm that is well-known because of its sensitivity to the modifications in the values of its constraints. Particle Swarm Optimization (PSO) is a category of heuristic algorithm motivated from the nature social activities of birds. The most important potential of PSO is its speedy convergence, when compared against other global optimization approaches[11]. The main contribution of the CFP-Growth++ and MPSO-LSSVM based association rule mining work is as follows:

1. The CFP-growth algorithm utilizes for finding frequent patterns.
2. Constructing the Minimum Item Support (MIS)-tree, CFP-array and producing frequent patterns from the MIS-tree.
3. The CFP-growth algorithm utilizes for finding frequent patterns where constructing the Minimum Item Support (MIS)-tree, CFP-array and producing frequent patterns from the MIS-tree.
4. From the set of frequent item sets found, create all the association rules that have a confidence exceeding the minimum confidence.
5. Enhanced outlier detection method for determining the outlier degree from association rules for outlier detection.
6. Association rules are mined using MPSO-LSSVM classification based association rule mining algorithm.

The rest of this paper is organized as follows. Section 2 explains the proposed CFP-Growth++ with outlier detection scheme. Section 3 introduces the proposed hybrid algorithm. Section 4 clearly discusses the experimental results. At last, a section gives brief conclusion and further research.

2. Proposed Work

The Type 2 Diabetes (TTD) is a lot of accepted kind of diabetes and reports for 90-95% of all diabetes. Detection of TTD from assorted factors or symptoms became an affair which is not free from false presumptions accompanied by unreliable effects. Along with this circumstance, the data mining could be acclimated, advice us in ability analysis from data. This work using Support Vector Machine (SVM) in the data mining action access information from actual data of patient medical records. It presents a decision-making support through association rule mining based on CFP-Growth++.

The Figure 1 illustrates the block diagram of the proposed system. Initially, the Type-II DM patient data is given as input. This may contain unwanted and empty data this significantly reduce the detection accuracy. In next step, pre-processing is done to remove the noisy data. This is done through the Chi Merge discretization method. After that CFP-growth algorithm utilizes for finding frequent patterns of the input dataset. The resultant association rules from CFP-Growth++ are discovered using MPSO-LS-SVM Classification algorithm, as a result the accurate result of association rules are generated. Here the enhanced outlier detection technique is used for determining the outlier degree from association rules for outlier detection. As a result the outlier transaction detected by enhancing an association classification (MPSO-LSSVM) approach using CFP-Growth++ algorithm.

2.1 Datasets Used

In this section, the input patient dataset is obtained from Pima Indians Diabetes Data Set. This contains the reports of both men and women at least from 10 years old. Here TRUE and FALSE values are assigned for the variable where TRUE denotes a positive test for TTD and FALSE denotes a negative test. In this input dataset 185 cases are there, where 133 cases in class TRUE and 52 cases in class FALSE. There are ten clinical attributes are shown in Table 1.

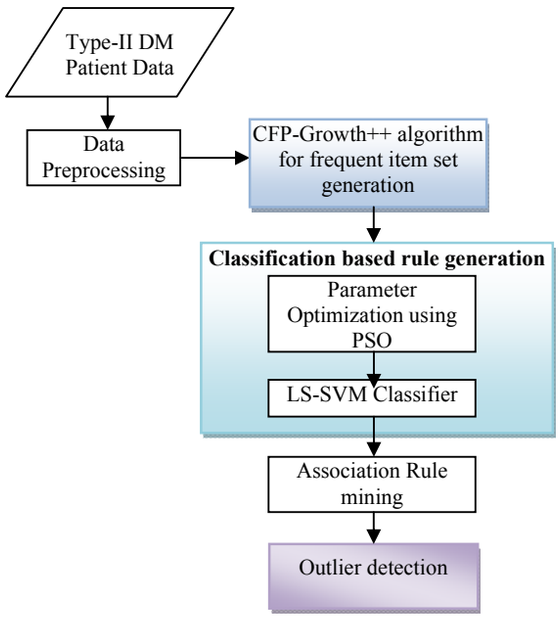

Figure 1. Block diagram for enhancing an association classification using CFP-Growth++ algorithm.

2.2 Data Processing

Discretization is the procedure of translating real data attributes into a naturally small digit of finite values. The difference in the original data is preserved in the discretized dataset. Discretization is an essential originator to using association mining algorithms. Generally data attributes are moreover discrete or categorical. Though categorical attributes are discrete, numerical attributes are moreover discrete or continuous. In this method each interval can be treated as one value of a discrete attribute, since it partitions an attribute's values into a number of intervals (min_i ... max_i). The preference of the intervals could be detected through an area specialized expert or with the advice of an automated action, which formulates the discretization process will be quick. In this work Chi Merge is utilized as a discretization method.

Chi Merge is one kind of discretize methods intended for discretizing the data and it utilizes Chi Square statistics that is efficient on the attributes of a record. The neighboring pairs of values are evaluated to discover the match among the data using chi square analysis. If the data are comparable then they are reserved in same interval and if not then they are put in different intervals.

Step 1: Calculate the (x^2) value for each A_{adj} //pair of adjacent intervals
Step 2: Combine A_{adj} with lowest (x^2) value
Step 3: Do again step 1 and step 2 until (x^2) values of all adjacent pairs exceeds a threshold

Table 1. Type-2 diabetes attributes

1. Body mass (thin, medium, overweight) (weight in kg/(height in m)^2)
2. Blood pressure (< 140/90, ≥ 140/90)
3. Hyperlipidemia (true, false)
4. Fasting blood sugar (FBS) (< 126 mg/dl, ≥ 126 mg/dl)
5. Instant blood sugar (< 200 mg/dl, ≥ 200 mg/dl)
6. Diabetes Gest history (true, false)
7. Plasma insulin (high, low)
8. Number of times pregnant
9. Diastolic blood pressure (mm Hg)
10. Triceps skin fold thickness (mm)
11. 2-Hour serum insulin (mu U/ml)
12. Diabetes pedigree function
13. Age (years)
14. Class variable (0 or 1)

Where, $x^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$, o_{ij} is the observed frequency of interval i for class j and e_{ij} is the expected frequency $\frac{(R_i * C_j)}{N}$.

2.3 Frequent Pattern Generation Using CFP-Growth++ Algorithm

An association rule is an inference of the form: X⇒Y, where X⊆I, Y⊆I, and X∩Y=∅ that of each association rule has a support and a confidence value. The support value of association rule X⇒Y is the proportion of transactions in DB that having both X and Y. The projected work is separated into two parts as specified below:

1. Discover all frequent item sets.
2. From the set of frequent item sets found, create all the association rules that have a confidence exceeding the minimum confidence.

2.3.1 CFP-Growth++ Algorithm

In this work, the CFP-growth algorithm utilizes for finding frequent patterns. The suggested algorithm called as Complete Frequent Pattern-growth++ (CFP-growth++) which involves three processes such as constructing the Minimum Item Support (MIS)-tree, CFP-array and producing frequent patterns from the MIS-tree. The CFP-growth++ determines the perceptions particularly least minimum support and uncommon child node pruning for building the MIS-tree for the purpose that the volume of the consequential MIS-tree probably less than or comparable to the MIS- tree generated by CFP-growth advance and CFP-array is dropping memory exploitation. For all item, the CFP-growth++ hypothesizes that user indicates the MIS values earlier to its accomplishment. Accordingly, the repeated patterns are constructed with a solitary scan on the dataset by the use of the priori information, in particular MIS values of the items. The three most important procedure of CFP-growth++ is given in detail in the following sub-section:

Least Minimum Support: The frequent patterns obtained by the use of multiple minsup values with the arranged closure property, explicitly, the entire supersets relating to the item having least MIS value is supposed to be common in a frequent pattern. Therefore, frequent item indicates the item having the least MIS value in each frequent pattern. In view of that, among all the common items, every frequent pattern will have support bigger than or almost equal to least MIS value. Thus, in case, when the support value is lesser than the least MIS value of the common item, subsequently all the items eradicated where no frequent pattern will be disregarded. This perception is concerned as Least Minimum Support (LMS) that referring the lowest MIS value among all the common items.

Constructing MIS-tree: The CFP- growth++ algorithm acquires the input constraints of the items, which is specifically, transaction dataset indicated as Trans, Item set denoted as I and minimum item support values signified as MIS. The CFP++-growth builds a preliminary MIS-tree with these input constraints that is associated with MIS-tree constructed by CFP-growth. Next, tree-pruning is accomplished to eradicate the uncommon items from the item-header table and MIS-tree, beginning from the last item in the item-header table whose item having least MIS value. At once solitary item is pruned in item-header table then progress to next item and perform tree pruning. However, it ends tree-pruning procedure when the frequent item is confronted where the MIS value of this common item is the LMS significance. In the end, prune the uncommon child nodes in the MIS-tree following tree merging process is terminated and the resulting MIS-tree is the condensed MIS-tree.

Mining Frequent Patterns from Mis-Tree: The practice of mining the MIS-tree in CFP-growth++ is practically comparable like mining the MIS-tree in CFP-growth. On the other hand, the difference among CFP-Growth++ and CFP-Growth approaches is that for each item in the header of the tree, the CFP++-growth approach validates whether the suffix item in the header of the Tree is a frequent item, before building conditional pattern base and conditional MIS-tree. In accordance with this condition, when a suffix item is not a common pattern then the construction of conditional pattern base and conditional MIS-tree are neglected. This suitable point is exploited to choose the item having least MIS value that is supposed to be a frequent item in each frequent pattern. In building the conditional pattern support for a suffix item, the suffix item implies the item which has least MIS value. As a result, when an uncommon item is the suffix item subsequently the patterns will be infrequent where it implies the item having least MIS value. Consequently, it is not indispensable to build conditional pattern base for an infrequent suffix item.

Infrequent child node pruning: As discussed above, the CFP-growth++ method skips the formation of conditional pattern supports for the infrequent suffix items. In view of the fact that its prefix paths, in particular conditional pattern bases are not exploited, the child nodes will be a member of uncommon items whose have no influence in the compact MIS-tree. Accordingly, when the child nodes prune interrelated with infrequent items, still the consequential MIS-tree will secure the transaction details relating to frequent patterns. So, infrequent child node pruning practice is implemented in the MIS-tree in order that every branch terminates with the node of a frequent item. Specifically, pruning should be achieved merely on the child nodes associated with infrequent items.

Following are some of the rules generated using CFP-growth++ mining that helps to discover basis of diabetes.

Rule 1: Urine Alb. < 300}{Heart problem is absent}{Ceratine is Negative}{TG < 250}{Uric Acid is Absent}{LDL is Low}–>{T2DM is present} 96.66%

Rule 2: {Ceratine is Negative}{TG < 250}{SGPT has no value}{HDL <35}{LDL is Low}––>{T2DM is present} 96.66%

Rule 3: {Urine Alb. < 300}{Ceratine is Negative}{TG < 250}{SGPT has no value}{HDL < 35}{LDL is Low}–> {T2DM is present} 96.66%

Rule 4: {Heart problem is absent}{Ceratine is Negative}{TG < 250}{SGPT has no value}{HDL < 35}{LDL is Low}–> {T2DM is present} 96.66%

Rule 5: {Urine Alb. < 300}{Heart problem is absent}{Ceratine is Negative}{TG < 250}{SGPT has no value}{HDL < 35}{LDL is Low}––> {T2DM is present} 96.66%

2.3.2 Enhanced Outlier Detection Method

In this section, the technique for determining the outlier degree from association rules for outlier detection is summarized. An association rule X $*\rightarrow$ Y with a high confidence denotes that when X happens, then Y happens with high probability. To be exact, when X arises in a transaction subsequently all items contained in Y should also arise in the transaction. Abuse of the rule is a warning of the outlier transactions. The outlier degree utilized in the proposed work is depends on associative closure property of a transaction and is illustrated as follows. Let 't' be the set of transactions ($t^{i+1} \subseteq t'$) with high confidence rules (R), then its associative closure t' is defined as below:

$$t^0 = t$$

$$t^{i+1} = t^i \cup \{e/e \in Y \text{ for every } X \subseteq t^i \text{ for all } X ? \rightarrow Y \in R\} \quad (1)$$

$$t' = t^{\infty}$$

When the item set t has unobserved rules, consistent with this definition, for a transaction t the cardinality of item set t^{i+1} increases. If t^i has no ignored rules then t^{i+1} converge turn out to be associative closure t′. The associative closure t′ is exclusive for each t. In a transaction, t, the variation between t and its ideal form t′ becomes bigger when the number of items with strong dependency is small. If t has less ignored rules, then t′ is comparable to t and decided is not an outlier. Here, outlier degree is computed as $O_d(t) = \frac{|t' - t|}{|t'|}$ when t′ =t,∅and $O_d(t) = 0,1$. Accordingly, O_d is constantly a value between 0 and 1, with lower bound of $O_d(t)$ is 0 and upper bound is $O_d(t) < 1$. Rules with 100% confidence are ignored from outlier checking, because they have items, have high confidence and cannot be unobservable rules for any other transactions. The outlier transactions can be known from the outlier degree. If $O_d(t) \geq \min_{(od)}$ where $\min_{(od)}$ is the minimum outlier detection and is user defined then the transaction 't' ($t \in T$) is supposed to be an outlier transaction.

3. Mining Association Rule Using MPSO-LSSVM

3.1 Modified Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a method motivated from the characters of normal social behavior, stimulating motions and interactions of insects, birds and fish; this is discussed in [12]. The major strength of character of PSO is its speedy convergence, when comparing against most successful global optimization approaches like Genetic Algorithms (GA), Simulated Annealing (SA), Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO). The main idea is to take care of the modifications in velocity. Overall, the main idea of PSO is as follows. Consider a particle in *d* dimension, it could revise its acceleration and position with the help of (2) and, (3). In which r_1 and r_2 represents two random numbers in the range [0, 1], V_{id} represents the momentum, ω_{id} denotes the inertia

weight, C_1 denotes the cognitive learning parameter and C_2 represents the social collaboration parameter. The location of the ith particle $X_{id} = (x_{i1}, x_{i2}, \ldots x_{id})$ is, $P_i = (p_{i1}, p_{i2}, \ldots p_{id})$ indicates the finest earlier position that is the location with the maximum fitness value.

$$V_{id} = \omega_{id} V_{id} + C_1 r_1 (p_{id} - X_{id}) + C_2 r_2 (p_{gd} - X_{id}) \quad (2)$$

$$X_{id} = X_{id} + V_{id} \quad (3)$$

A crucial role played by inertia weight in the practice of offering steadiness between searching and exploitation. It determines the contribution rate of a particles earlier velocity to its recent velocity at the present time. Several categories of inertia weights were discovered like constant, random, adaptive inertia weight and a lot of supplementary types. In this research work, an enhanced version of PSO is developed. It possibly will revise its velocity and position with the help of (4) and (5) for the i th particle in d dimension;

$$V_{id} = \lambda[\omega_{id} V_{id} + C_r r_1 (p_{id} - X_{id}) + C_2 r_2 (p_{gd} - X_{id})] \quad (4)$$

$$X_{id} = X_{id} + (\omega V_{id}) \quad (5)$$

Where λ represents a convergence factor that can be calculated by means of (6)

$$\lambda = \frac{2}{\left|2 - C - \sqrt{C^2 - 4C}\right|} \quad (6)$$

Where $C = C_1 + C_2$

In this algorithm ω_{id} is estimated with the assistance of (7), in which t represents iterate of overall iterations and T_{max} indicates the maximum number of iterations. The ω will be lessened with the increasing of t linearly from 0.9 to 0.4[6];

$$\omega_{id} = 0.9 - \frac{t}{T_{max}} * 0.5 \quad (7)$$

3.2 Least Squares Support Vector Machine

Least Squares-Support Vector Machine (LS-SVM) classifier is a kind of model of Support Vector Machine (SVM)[13]. With the support of this classifier, it is uncomplicated to find out the solution by solving a collection of linear equations more willingly than a convex quadratic programming difficulty for standard SVMs. Finding out a best possible hyper plane is the major objective of LS-SVM that segments diverse classes. This objective can be accomplished with the help of maximum Euclidean distance to the nearby location. The LS-SVM classifier plots the input vectors into a peak dimensional characteristic space for non-separable data. Subsequently, the LS-SVM classifier determines a best possible among the hyper-plane in this high dimensional characteristic space[14].

For a particular training dataset of N points $\{x_k, y_k\}_{k=1}^{N}$ with input data $x_k \in R^n$ and output $y_k \in R$, the subsequent optimization setback measured in original weight space:

$$\min J(w,b)_{w,b,e} = \frac{1}{2} W^T w + \frac{1}{2} \gamma \sum_{k=1}^{N} e_k^2 \quad (8)$$

$$\text{Such that } y_k - (w^T \varphi x_k + b) = e_k, k = 1, 2 \ldots N \quad (9)$$

Where, γ represents regularization factor, e_k indicates the difference among the real output and the preferred output y_k, and $\varphi(.)$ represents a nonlinear function mapping the data points into a high dimensional Hilbert space. Furthermore, the dot product in the high-dimensional space is equivalent with a positive definite kernel function $K(x_i, x_j) = \varphi(x_i)^T (x_j)$. A linear classifier in the fresh space acquires the given structure, in which w represents the weight vector and $b \in R$ which is regarded as the bias expression in primitive weight space.

$$y(x) = \text{sign}(w.\ \varphi(x) + b) \quad (10)$$

The twofold space of this original space is found out with the help of the Lagrangian function given in, (11)

$$L(w, e, \propto) = J(w, e) - \sum_{k=1}^{N} \propto_k (w^T \varphi(x_k) + e_k - y_k) \quad (11)$$

Where L (w,e,$\propto$) indicates the Lagrangian multipliers and are known as Support Vectors. The best possible solution for objective function in, (11) must guarantee the Karush-Kuhn Tucker (KKT) stipulations (12) as given below:

$$\frac{\delta L}{\delta w} = 0 \rightarrow w = \sum_{k=1}^{N} \alpha_k y_k \varphi(x_k) \quad (12)$$

$$\frac{\delta L}{\delta w} = 0 \rightarrow \propto_k = \gamma e_k, k = 1, 2 \ldots N$$

$$\frac{\delta L}{\delta w} = 0 \rightarrow w^T \varphi(x_k) + e_k - y_k = 0, k = 1, 2 \ldots N$$

The linear system in, (13) will be the outcome following the removal of w and e which produces the Support Vector $\propto_k'$

$$\left(K + \frac{1}{\sigma}\right) \alpha = y \quad (13)$$

Where $y = [y_1, y_2, \ldots y_N]^T, \propto = [\propto_1, \propto_2, \ldots \propto_N]^T$ and $K \in R^{N \times N}$ represents the kernel matrix. The resultant LS-SVM model for function evaluation is as given in (14) in which $K(x, x_k)$ is the kernel function;

$$y(x) = \sum_{k=1}^{N} \propto_k K(x, x_k) \quad (14)$$

LS-SVM is implemented using Ra-dial Basis Function (RBF), (15).

$$K(x, x_k) = \exp\left(-\frac{|x - x_k|^2}{\sigma^2} \right) \quad (15)$$

3.3 MPSO-LSSVM Approach for Rule Discovery

Association rule discovery and main mining process that intends at revealing all frequent patterns among transactions consist of data attributes or items. The major difficulty occurs because of the fact that there is various number of transaction data that are take into account for generating the association rules using SVM Classification algorithm. The high confidence value based classifier utilized for classifying the specified input data to that medical domain. Association rules are mined from CFP-Growth++ and using LS-SVM Classification algorithm the accurate result of association rules are generated. A test medical data is labeled into one of the predefined classes labels depending on the high confidence value. With the intention of reducing the irrelevant association rules, a hybrid classification algorithm developed that incorporates Modified-PSO algorithm as a parameters optimization approach and LS-SVM for classification that includes two main phases, Training phase and a Testing phase.. The algorithm for Type-2 DM diagnosis and treatment also includes two main phases, Parameters Optimization and Classification. Modified-PSO technique is exploited as parameters optimization method aimed to advance the setting of the parameter values of LS-SVM. Consequently, surpassing its sensitivity to the parameter values modifications.

The main intention of exploiting Modified-PSO for parameters optimization phase is to decide the best possible values for the parameters of the LS-SVM classifier, for instance, the regularization factor (σ) and Gaussian Kernel function (γ). Furthermore, next stage categorizes the Type-2 DM patients with the optimized parameters into classes by means of LS-SVM. The step by step algorithm is given below.

Input: Frequent pattern item set

Output: Best possible parameters values //from the result Association rule creation is done

Step 1: Give the Frequent pattern item set of *n* data points,

Step 2: Generate random weights for all input data point.

Step 3: Begin the bias term *b* and the error *e* for each point arbitrarily.

Step 4: Find out the best possible values γ and σ with the assistance of MPSO.

Step 5: Realize best possible values for the objective function with the support of

$$\min J(w, b)_{w,b,e} = \frac{1}{2}\gamma \sum_{k=1}^{N} e_k^2$$

Step 6: Work out number of support vectors ($\propto$) by means of $\left(K + \frac{I}{\sigma} \right) = y$

Step 7: Categorize any new point by $y(x) = sign(w.\varphi(x) + b)$ with the support of RBF kernel function $K(x, x_k) = \exp\left(-\frac{|x - x_k|^2}{\sigma^2} \right)$

Step 8: Proceed until stopping criteria is attained, generally until reach the maximum number of iterations.

3.3.1 Association Rule Generation

In this section, the rule schemas are consists four operators such as conforming, pruning, and unexpected and exceptions[15]. The rules are generated from CFP-Growth++ algorithm which is constructed and validated. Rule schemas are defined for each and every operator as follows: Let Cof.(Rsc) be the conforming rule schema, Cof. (Rsc): X and Y → Z. Where, X, Y and Z are generalized concepts from our algorithm. They are denoted by the symbol $f_{Rsc}()$. The Rule Schema's which are to be defined for each operator from our algorithm is Conforming Rule Schemas: For Example:

Rsc: f_{Rsc}(Attributes) → f_{Rsc}(Values)

f_{Rsc}(Attributes) → {Gender, Body mass, FBS}

f_{Rsc}(Values) → {female, overweight, 126 mg/dl, yes/no}

R1: female, overweight, $126 \frac{mg}{dl} \rightarrow$ no

Number of transaction, N is calculates support (in percent) for each association rule is merely a ratio between support count and the number of transaction. Confidence

is computed simply by taking ratio of support counts of the union of the dependent variable to the support count of dependent variable. The next step to compute association rules of Type-2 DM diagnosis analysis is to concern two threshold criteria: minimum support and minimum confidence. Thus suppose to set two thresholds into cells for example supposes set 50% minimum support and 75% minimum confidence. Minimum Support = 50%, Minimum Confidence = 75%. Unusual threshold value will construct broader or stricter rules. As revealed above, an advantage of this proposal is that it presents understandable rules and it can be adapted and applied to each specific difficulty.

4. Experimental Result and Discussion

The performance of the proposed CFP-Growth++ algorithm with MPSO-LSSVM classification is described in this section. The data source of experiments is explained in Section 2. The CFP-Growth++ algorithm discovers frequent item sets and shows much greater efficiency than the Enhanced FP-Growth algorithm. The algorithm CFP-Growth with MPSO-LSSVM is reportedly working efficiently and in many cases, it's much faster than enhanced FP Growth with ACO. The results are found to be more interesting than association rules mined by CFP-Growth++ although they are the subsets of item sets mined by enhanced FP-Growth.

As shown in Figure 2 the proposed CFP-growth++ with MPSO-SVM algorithm is found better in both comparisons and also requires low computational time than other algorithms for same categorical attributes.

After comparing the performance of CFP-growth++ with MPSO-SVM algorithm (CFP-Growth++-MPSO-LSSVM) to the FP-Growth with ACO (FP-Growth-ACO) algorithm, it is clear that CFP-growth++ with MPSO-SVM algorithm is approximate 83% more accurate after using same data set. CFP-growth++ with MPSO-SVM algorithm generating time is less than 12 sec but FP-Growth with ACO algorithm is more than 46 sec. Figure 3 shows comparing result of both algorithms. As discussed earlier, an improvement of this approach is that it offers clear rules and permits a context-free grammar to be adapted and implemented to each definite problem.

The outcome when considering the outlier detection rate confirms that the CFP-growth++ with MPSO-SVM system is an advanced techniques of the base model in discovering outlier transactions shown in Figure 4. Additionally, it can also be proved that the performance of outlier detection increases with the quantity of outlier transaction exist in the transaction database.

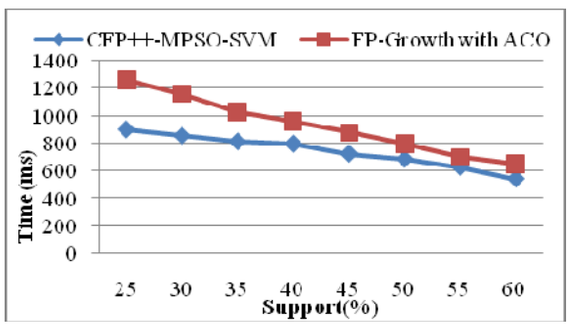

Figure 2. Performance analysis of both algorithms.

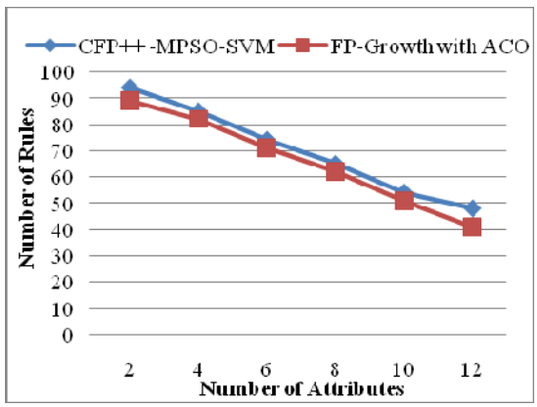

Figure 3. Generated association rules

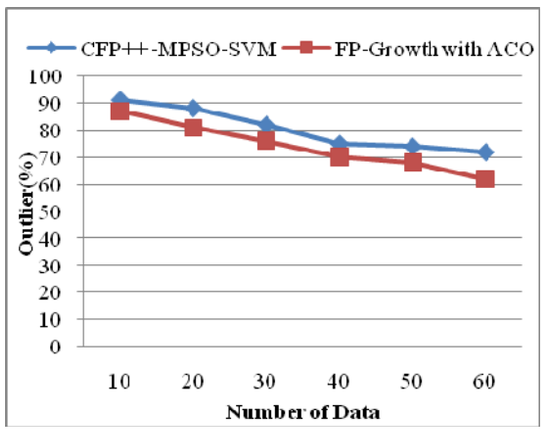

Figure 4. Outlier detection.

5. Conclusion

LS-SVM-based algorithm for mining association rules from Type-2 DM dataset is presented in this work. This algorithm based on LS-SVM Classification that permits both categorical and numerical attributes to be classified. In this work, an efficient algorithm called CFP-Growth++ proposed by using the notions as least minimum support and infrequent child node pruning. With the intention that the size of the resultant MIS-tree possibly will be less than or equivalent to the MIS-tree constructed by CFP-growth and enhanced FP-Growth approach. One of the most notable aspects of this algorithm is the exercise of some logical operators which permit frequent items to be attained in datasets where there are not many frequent patterns. This mining of association rules permits to attain understandable close relations between items, while these rules are more functional and reasonable. The enhanced outlier detection method identifies outlier transaction by enhancing an association classification approach using CFP-Growth++. The various experimental results projected to demonstrate that the proposed model are scalable and competent in terms of outlier detection and classification and can be utilized by data mining techniques for accurate and fast association rule discovery. From the experimental results CFP-Growth++-MPSO-SVM algorithm is found better in both comparisons and also requires low computational time than other algorithms. The memory requirements of CFP-growth++ approach will not at all exceed those of CFP-growth and Enhanced FP-Growth approach. Generally, it requires relatively less memory. In future the work can be aimed to solve this problem with that swarm optimization based ABC algorithm can be utilized as a pre-processing method, and its effect on the classification rate can be investigated.

6. References

1. Alberti KG, Zimmet PZ. Definition, diagnosis and classification of diabetes mellitus and its complications. part 1: diagnosis and classification of diabetes mellitus. Provisional report of a who consultation. Diabet Med. 1998; 15(7):539–53.
2. Keech A, Simes RJ, Barter P, Best J, Scott R, Taskinen M-R, Forder P, Pillai A, Davis T, Glasziou P, et al. Effects of long-term fenofibrate therapy on cardiovascular events in 9795 people with type 2 diabetes mellitus (the field study): randomized controlled trial. Lancet. 2005; 366(9500):1849–61.
3. Tuomilehto J, Lindstrom J, Eriks-son JG, Valle TT, Hamalainen H, Ilanne-Parikka P, Keinanen-Kiukaanniemi S, Laakso M, Louheranta A, Rastas M, et al. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. New Engl J Med. 2001; 344(18):1343–50.
4. Han J, Kamber M. Data mining: concepts and techniques. 2nd ed. San Fransisco, CA: Morgan Kaufmann; 2006.
5. Liu B, Hsu W, Ma Y. Integrating classification and association rule mining. Proceedings International Conference Knowledge Discovery Data Mining; New York: 1998. p. 80–6.
6. Rak R, Reformat LKM. A tree-projection-based algorithm for multi-label recurrent-item associative classification rule generation. Data Knowl Eng. 2008; 64(1):171–97.
7. Zhang C, Zhang S. Association rule mining: models and algorithms series (Lecture Notes Computer Science Series 2307). Berlin, Germany: SpringerVerlag; 2002.
8. Cherkasski V, Mulier F. Learning from data:concepts, theory, and methods. New York: WileyInterscience; 1998.
9. Tan P. Steinbach M, Kumar V. Introduction to data mining. Boston, MA: Addison-Wesley/Longman; 2005.
10. Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett, 1999; 9(3):293–300.
11. Blondin J. Particle swarm optimization: a tutorial. Availaible from: http://cs. armstrong. edu/saad/csci8100/pso tutorial. pdf, 2009.
12. Shi Y, et al. Particle swarm optimization: developments, applications and resources. Proceedings of the 2001 Congress on Evolutionary Computation 2001. 2001; 1:81–6.
13. Ye J, Xiong T. Svm versus least squares svm. International Conference on Artificial Intelligence and Statistics; 2007. p. 644–51.
14. Shao X, Wu K, Liao B. Single directional smoalgorithm for least squares support vector machines. Comput Intell Neurosci. 2013.
15. Karthikeyan T, Ragavan R, Vembandasamy K. Hierarchical K-means clustering algorithm for an E-care of diabetes mellitus. Int J Adv Res Comput Sci Software Eng. 2011 Dec; 3(12):653–60.
16. Jabbar MA, Deekshatulu BL, Chandra P. Heart disease prediction using lazy associative classification. International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s); 2013. p. 40–6.
17. Tian X, Hu G, Li J. A new fuzzy associative classification based on axiomatic fuzzy set theory. 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD); Yantai, Shandong: IEEE; 2010. p. 1152–6.

18. Pach F, Gyenesei A, Abonyi J. Compact fuzzy association rule based classifier. Expert Syst Appl. 2008; 34(4):2406–16.
19. Chen G, Liu H, et al. A new approach to classification based on association rule mining. Decis Support Syst. 2006; 42:674–89.
20. Karthikeyan T, Vembandasamy K. A refined continuous ant colony optimization based FP-growth association rule technique on type 2 diabetes. IRECOS. 2014; 9(8):1476–83.
21. Bouzouita I, Elloumi S. Integrated generic association rule based classifier. 18th International Workshop on Database and Expert Systems Applications, 2007. DEXA'07; IEEE; 2007 Sep. p. 514–8.
22. Buddeewong S, Kreesuradej W. A new association rule-based text classifier algorithm. 17th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 05; IEEE; 2005 Nov. p. 2.
23. Sabeti M, Sadreddini MH, Nezhad JT. EEG signal classification using an association rule-based classifier. IEEE International Conference on Signal Processing and Communications, ICSPC 2007; IEEE; 2007 Nov. p. 620–3.
24. Depaire B, Vanhoof K, Wets G. ARUBAS: an association rule based similarity framework for associative classifiers. IEEE International Conference on Data Mining Workshops, ICDMW'08; IEEE; 2008. p. 692–9.
25. Fazzolari M, Alcala R, Nojima Y, Ishibuchi H, Herrera F. Improving a fuzzy association rule-based classification model by granularity learning based on heuristic measures over multiple granularities. 2013 IEEE International Workshop on Genetic and Evolutionary Fuzzy Systems (GEFS); IEEE; 2013 Apr. p. 44–51.