ISSN (online) : 2277-5390
ISSN (Print): 2277-5382

The Multiple time series clinical data processing with modified artificial bee colony algorithm and artificial neural network

Priyanga M*[1], Dr. K. Sasi Kala Rani [2], Pavithra M[3], Yamunadevi S[4]

[1,3,4] *Department of Computer Science and Engineering, Hindusthan Institute of Technology,Coimbatore, Tamilnadu*
[2]*Head of the Dept, Department of Computer Science and Engineering, Hindusthan Institute of Technology,Coimbatore, Tamilnadu*
[1] priyangame2016@gmail.com [2] sasikalarani2016@gmail.com [3,4] pavithrame@gmail.com

Abstract

Objectives: The main objective of this research is to discover patient acuity or severity of illness for immediate practical use for clinicians by evaluating the use of multivariate time series modelling along with multiple models.
Methods/Statistical analysis: As large-scale multivariate time series data become increasingly common in application domains, such as health care and traffic analysis, researchers are challenged to build efficient tools to analyze it and provide useful insights. In many situations, analyzing a time-series in isolation is reasonable. And also this scenario is used to increase the prediction accuracy and reducing the time complexity using optimization algorithm.
Findings: The various research works has been analyzed and evaluated. From the analysis, the multiple measurements support vector machine (MMSVM), multiple measurements random forest regression (MMRF) and improved particle swarm optimization (IPSO) algorithm, modified artificial bee colony algorithm (MABCA) to solve the multiple time series problems by maximizing the optimal feature information which found to be superior for higher performance in terms of accuracy, precision and recall. The proposed MABCA with transductive support vector machine (TSVM) and artificial neural network (ANN) is used to improve the classification performance.
Application/Improvements: The findings of this work prove that the graph search based method provides better result than other approaches.

Keyword: Data mining, multiple measurements, support vector machine (SVM), particle swarm optimization (PSO), modified artificial bee colony algorithm (MABCA) and artificial neural network (ANN).

1. Introduction

Data mining is the process of searching hidden patterns in the huge amount of database. It also mine the patterns form data. The patterns and correlation between patterns can be found by scans via a huge volume of data. To make a data into information data mining plays an important role. Data mining requires the use of data analysis tool to determine previously unknown, valid patterns and relationships in huge volume data. Such kind of tool can enclose statistical model, mathematical algorithms and machine learning methods. Thus, data mining consists of more than gathering and running data, it also contains analysis and prediction.

Data mining tools includes variety of tasks. The major functionality of the task is analyzing the data and generates the results. It becomes an increasingly common in both private and public sectors. Associations used the data mining concept and tools for surveying the customer information, avoid fraud and misuse, and help in medical research. Data mining process includes three stages as follows. The initial exploration, model constructing or pattern identification with verification or validation and deployment.

The data mining process is applied in business transactions, medical information data, scientific data, satellite data and software engineering data. The huge volume of data is stored in files, databases and repositories, it is progressively significant in recent years. To increase potentially means then analysis and interpretation of information data will be extracting the interesting knowledge which is used to proper decision making [1] [2].

The data mining comprises of various steps they are data cleaning, data integration, data selection, data transformation, data mining, pattern assessment and knowledge demonstration.Data cleaning and knowledge integr ation has been enforced along during a pre-processing step to create an information warehouse. Data selection and transformation could be combined together to discover the knowledge representation. Data mining techniques are such as support vector machine (SVM), regression algorithms and optimization algorithms used to predict the decision from the specified dataset efficiently.

In[3] presented a merging large databases acquired from different sources with heterogeneous representations of information has become an increasingly important and difficult problem for many organizations. Instances of this problem appearing in the literature have been called record linkage, the semantic integration problem or the instance identification problem and more recently the data cleansing problem regarded as a crucial first step in a Knowledge Discovery in Databases (KDD) process. Business organizations call this problem the merge/purge problem.The problem of merging multiple databases of information about common entities is frequently encountered in KDD and decision support applications in large commercial and government organizations. The system provides a rule programming module that is easy to program and quite good at finding duplicates especially in an environment with massive amounts of data.

In [4] presented a challenge of handling an ever increasing amount of data. In order to respond quickly to changes and make logical decisions the management needs rapid access to information in order to research the past and identify relevant trends. This information is usually kept in very large operational databases and the easiest way to gain access to this data and facilitate strategic decision making is to set up a data warehouse. Data mining techniques can then be used to find optimal clustering or interesting irregularities in the data warehouse because these techniques are able to zoom in on interesting subparts of the warehouse.

In [5] presents a simple and fast comparison method, TI Similarity, which reduces the time for each comparison. A new detection method RAR is proposed to further reduce the number of comparisons. With RAR and TI similarity, this approach for cleansing large databases is composed of two processes: Filtering process and Pruning process. In filtering process, a fast scan on the database is carried out with RAR and TI-Similarity. In pruning process, the duplicate result from the filtering process is pruned to eliminate the false positives using more trust worthy comparison methods. Related works such as SNM, Clustering SNM, Multi-pass SNM, Equation theory, Record Similarity (RS) and Edit Distance are discussed along with comparisons between methods. Existing data cleansing methods are costly and will take very long time for cleansing large databases. Large proportion of time in data cleansing is spent on the comparisons of records. TI similarity is a simple and fast comparison method. It computes the field's weights and filed similarity and the degree of similarity for records. RAR (Reduction using Anchor Record) is a fast detection method. It is also "Sorting and the merging" based and thus can be summarized in the three phases: Create key, Sort data and Merge. The previous two phases are the same as those in SNM. RAR uses the D-rule and ND-rule to reduce unnecessary comparisons.

In [6] presented a data integration is that the downside of combining data residing at different sources, and providing the user with a unified view of these data. The trouble of designing data integration systems is essential in contemporary real world programs, and is characterized by using some of problems which can be exciting from a theoretical point of view. This tutorial is targeted on some of those theoretical troubles, with unique emphasis on the following topics. The data integration systems are interested in this work is characterized by an architecture based on a set of sources and a global schema. The sources contain the real data, while the global schema provides a reconciled, integrated, and virtual view of the underlying sources. Modelling the relation between the sources and the worldwide schema is therefore a critical thing. For this purpose two basic approaches have been proposed. In the first approach the global schema is expressed in terms of the data sources which approach is called global-as-view. Whereas in the second approach, called local-as-view, requires the global schema to be specified independently from the sources, and the relationships between the global schema and the sources are established by defining every source as a view over the global schema.

In [7] presented an ontology which provides a distinctly dynamic and bendy map of the information contained within the information resources inside a domain. Due to the fact ontology enable real semantic integration across the data resources that they represent, it is possible no longer only to attract wider conclusions from the data but additionally to observe the data from several awesome perspectives applicable to the specific job being undertaken. The era of ontologies is an essential interest to allow semantic data integration. Efficiency in data integration can be achieved through ontologies. Thus the continual redevelopment risk of project-specific integration strategies can be reduced.

In [8] offered a data integration which affords a continual undertaking faced in applications that want to query throughout multiple independent and heterogeneous data sources. Data integration is essential in massive firms that personal a large number of data sources for producing data sets that may expand. The aim of data integration system is to provide uniform access to a hard and fast of heterogonous data sources and to free the user from the information about how data are structured on the sources consists of wrapping data sources and either loading the retrieved data into a data warehouse or returning it to the user. Wrapping a data source means getting data from somewhere and translating it into a common integrated format.

In [9] presented about a chronic asthmatic sufferers need to be constantly observed to prevent sudden attacks. In order to improve the efficiency and effectiveness of patient monitoring, can proposed in this paper a novel data mining mechanism for predicting attacks of chronic diseases by considering of both bio-signals of patients and environmental factors. It proposed two data mining methods, namely Pattern Based Decision Tree (PBDT) and Pattern Based Class-Association Rule (PBCAR). Both methods integrate the concepts of sequential pattern mining to extract features of asthma attacks, and then build classifiers with the concepts of decision tree mining and rule-based method respectively. Besides the general clinical data of patients, can considered environmental factors, which are related to many chronic diseases.

In [10] presented the successful application of data mining in extraordinary visible fields like marketing, e-business and retail has led to its application in different industries and sectors. Among these sectors just discovering is healthcare. The healthcare environment is still 'information rich' but 'knowledge poor'. In the healthcare system there is a huge amount of data are available. But still there is a lack of effective analysis tools to discover hidden relationships and trends in data. This research paper intends to provide a survey of present day strategies of knowledge discovery in databases the use of data mining strategies which are in use in today's medical research mainly in coronary heart ailment prediction. Number of experiment has been conducted to compare the overall performance of predictive data mining technique at the equal dataset and the outcome well-known shows that decision tree outperforms and a while Bayesian classification is having similar accuracy as of decision tree but other predictive methods like kNN, neural networks, type primarily based on clustering are not performing well. Another conclusion is that the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction.

In [11] presented a leading disease about a breast cancer in women in developed countries. Earlier detection of breast cancer is the most effective way to reduce breast cancer deaths. Early prognosis calls for a correct and dependable analysis technique that lets in physicians to differentiate benign breast tumours from malignant ones without going for surgical biopsy. The main aim of these predictions is to assign patients to either a "benign" group that is noncancerous or a "malignant" group that is cancerous. The diagnosis problem is the long-term outlook for the disease for patients whose cancer has been surgically removed. In this hassle a patient is classified as a 'recur' if the disease is observed at some subsequent time to tumour excision and a affected person for whom cancer has now not recurred and might in no way to recur. The aim of these predictions is to deal with cases for which most cancers has not recurred (censored information) as well as case for which cancer has recurred at a particular time. Consequently, breast cancers diagnostic and prognostic issues are in particular within the scope of the extensively mentioned classification problems. These issues have attracted many researchers in computational intelligence, data mining, and statistics fields.

In [12] discussed multiple time series clinical data processing for classification with merging algorithm and statistical measures. Conventional data processing technique [13] and classification methods may cause medical data to disappear and decrease classification performance. To progress the precision of medical outcome classification using multiple measurements, a novel multiple time series data processing approach with merging algorithm is enhanced. Clinical information [14] from hepatocellular carcinoma (HCC) patients is utilized in this scenario. Their medical descriptions from a defined period are combined by using the improved merging algorithm, and statistical measures are also computed. Then, multiple measurement support vector machine (MMSVM) is utilized as a classification technique [15] to identify the HCC occurrences. A multiple measurements random forest regression (MMRF) is also utilized as a supplementary assessment [16] and prediction technique. To compare the data merging approach, the classification performance by using processed multiple measurements is compared to classification using single measurements [17]. Also the improved particle swarm optimization algorithm is introduced to increase the accuracy of classification results [18].

Thus this method has been identified as the better technique as the approach provides deep analysis of higher accuracy prediction for the specified datasets. The efficiency of this approach is better than the other techniques and also it provides pathway for further improvement.

ISSN (online) : 2277-5390
ISSN (Print): 2277-5382

2. Materials and Methods

2.1. Pre-processing

In this module, the pre processing technique is performed to obtain the more accurate classification results. Data cleaning is the process of discovering and correcting inaccurate records from the specified dataset. Used mainly in databases, the term refers to identifying incomplete, incorrect, inaccurate, irrelevant, etc. data integration is the process of combine the various information data from heterogeneous data sources but with semantic meaning. It is used to increase the classification accuracy results for the specific query. Data transformation is used to change the set of data values from the source data system to destination data system. By using the pre-processing method, the accuracy of classification performance is increased in terms of reduction of noise and missing values.

2.2. Feature selection

In this module we have to perform the feature selection process on the time series dataset. It is used to provide relevant feature for the training and testing process. To remove the redundant and irrelevant features, the feature selection based random forest is introduced. An ensemble classifier algorithm is enhanced which contains bagging and random feature selection methods. The frequency of a feature's appearance in the classification trees represents the importance of the feature. The library random forest is used to execute the random forest feature selection process. All the features are ranked according to the weight assigned to them by random forest.

2.3. Algorithm for Merging Multiple Features Based on Defined Time Periods

In this module, we have to merge the important features by using the merging algorithm more efficiently. Based on the algorithm 1 we evaluated the time series data. The algorithm is as follows.

Algorithm 1

Start
Read D_m= *m* days period
T_{event}= the time of specific event
R_B = all records before T_{event}, sorted by record date in descending order
F_B= all features inR_B
Initialize merged records array $M_m$based on *m* days period and F_B
FOR each record R_{Bk}in R_B, k=1,2…N
T_k = the time of $R_{B,k}$, recorded
i= $T_{event} - T_k / D_m$
M_{mi} = the *i*th merged record based on *m* days period
FOR each feature F_q in *RB k* (*q* = 1*, . . . , O*)
Set the value *Wq* of *Fq*in *Mm i*as the most recent value of *Fq* from all the *RB k* in *RB* and *i*th period
ENDFOR
ENDFOR
If statistical measures mode
FOR each period *i*in *Mm*
FOR each time-related feature *Ft* in *FB*
//time-related laboratory data in Supplementary Data 1
*Ft*_ M ax*i*= maximum of all the *Ft* in *RB* within period *i*
*Ft*_ M in*i*= minimum of all the *Ft* in *RB* within period *i*
Ft _A vg *i*= average of all the *Ft* in *RB* within period *i*
Ft _SD *i*= standard deviation of all the *Ft* in *RB* within period *i*
Ft _C or*i*= Pearson's correlation coefficient of all the *Ft* in *RB* within period *i*
Ft _S lp*i*= slope of trend line of all the *Ft* in *RB* within period *i*
Add *Ft* M ax*i*,*Ft* M in*i*, *Ft* A vg *i*, *Ft* SD *i*, *Ft* C or*i*, *Ft* S lp*i*as addition features into the *i*th merged record *Mm i*
ENDFOR
ENDFOR
OUTPUT *Mm*
END

The central idea of this merging algorithm is to choose only one value to stand for a feature in one period. Because the time of the target event, such as therapy for HCC, is set as the key time with regard to data processing, the value that is closer to event time could be more significant than others. Therefore, the most recent value is selected to represent a feature in a period, and therefore some valuable information in the original data might be omitted by the merging algorithm.

2.4. Calculation of statistical measure

In this module, statistical measure is calculated for describing the data distribution in each period. There is a probability that information in the original data, such as the tendency and feature distribution may disappear after data merging. To protect the information, maximum and minimum statistics are used in this scenario. Average is a method for deriving the central tendency of a feature space, and standard deviation is an extensively utilized measurement of variability. Pearson's correlation coefficient is showing, how the feature pair is strongly related within the range of -1 to +1.

2.5. Prediction model establishment

In this module, the data mining approaches are such as support vector machine (SVM) and random forest used for single and multiple measurements respectively. The SVM builds the classification model for a binary class and it uses nonlinear mapping to change the data into higher dimensional data. Along with a suitable nonlinear mapping, two classes are divided through a hyperplane. The library SVM is focused to execute the SVM prediction process. The kernel function with radial basis function is used for SVM model establishment. For multiple measurements, the prediction outcomes are decided through voting method where more features belonged to similar group and majority vote of class is considered as final prediction result.

Algorithm 2

BEGIN
*S*m = the test dataset selected from the merged records based on m days period
P V m = the patient list of test dataset *S*m
*R*m = the training dataset selected from the merged records based on *m* days period
PMm= the predictive model established based on selected features in *R*m , and
imported parameters
FOR each patient *Pi* in *P V m*
Initialize voting result of *Pi* ,*VR i*to zero
If the type of predictive model is classification
FOR each merged record *P Smi*of *Pi* in period *i*in *S*m
Rmi= prediction result of *P Smi*by using *PMm*
//recurrence = 1, non − recurrence = −1
VR i= *Rmi*+ *VR i*
ENDFOR
If *VR i*>= 0
Predict *Pi* as a positive case //recurrence
Else
Predict *Pi* as a negative case //non-recurrence
Else Ifthe type of predictive model is regression
VR i= the average of prediction result of all merged record of *Pi* in period *i*in
*S*m by using*PM*m
Predict *Pi* by *VR i*//regression result
ENDFOR
OUTPUT performance of *PMm* based on the prediction results
END

We evaluated the dataset by using MMSVM and MMRF algorithm efficiently.

ISSN (online) : 2277-5390
ISSN (Print): 2277-5382

2.6. IPSO classification

The set of rules proposed in this work is totally based on the particle swarm optimization approach. PSO is an optimization algorithm that optimizes a given solutions via applying mathematical rules and after computing the fitness of a current solutions modify their coordinates into the search space. PSO is initially delivered by Kennedy, Eberhard, and Shi as an optimization technique stimulated by the social behaviour of bird flocks and fish herds. PSO forms swarms by utilizing a certain number of solutions, called particles. Every such particle accomplished with position and velocity coordinates in the search space. The change of the particle position from iteration to iteration is represented as velocity. The amendment of the particle's position is determined by the simplest up to now best-known particle's position additionally from the simplest position within the overall swarm.

This is used to improve the speed of the process by using important and relevant information features in the dataset. It reduces the number of iterations by selecting the best solutions for time series dataset. The IPSO algorithm is as follows

Algorithm 3

```
UpdatePSO
{
Do
ForEach Particle in Swarm
For j = 0 to ParticleLength
 Partcle.Velocity[j] = W * Partcle.Velocity[j] + C1*R1*Particle.BestPosition[j] - Particle.Position[j] +C2*R2*BestParticle.Position[j] - Particle.Position[j]
EndFor
For j = 0 to ParticleLength
 Partcle.Position[j] += Partcle.Velocity[j]
EndFor
 CheckCandidate (Particle)
 If (Particle.BestInfoGain > BestParticle.BestInfoGain)
 BestParticle = Particle
 EndIf
EndForEach
 OldBestGain = NewBestGain
 NewBestGain = GetSwarmBestInformationGain
 While ( (OldBestGain - NewBestGain) > EPSILON )
 BestShapelet = BestParticle
}
IPSO
CheckCandidate(Particle)
{
 Distances ← Initialize
 ForEach TimeSeries in TrainDataSet_ClassA_And_ClassB
 Distance = MinDistance(Particle.Position, TimeSeries)
```

Distances ← Add(Distance)

EndForEach

Histogram = OrderDistances(Distances)

InforGain = CalculateInformationGain(Histogram)

If (InforGain > Particle.BestInfoGain)

Particle.BestInfoGain = InfoGain

Particle.BestPosition = Particle.Position

EndIf

In this proposed work, effective summarisation process is achieved through improved PSO approach. Particle swarm optimization (PSO) is a computational algorithm that optimizes a problem by iteratively trying to progress a candidate solution along with regard to a given measure of quality. The c1 and c2 are cognitive parameters, r1 and r2 are random parameters. It is used to choose the best solutions from the multiple time series data. PSO optimizes a problem by having a population of candidate solutions, here dubbed particles, and moving these particles around in the search-space according to simple mathematical formulae over the particle's position and velocity. Each particle's movement is influenced by its local best known position but, is also guided toward the best known positions in the search-space, which are updated as better positions are found by other particles. This is expected to move the swarm toward the best solutions.

In the proposed system, we introduced improved PSO algorithm to increase the classification accuracy. For the given input datasets, the similarities of multiple features are extracted optimally by using PSO parameters. The main aim of the PSO algorithm is to select the potential and relevant features by generating best fitness function value. Also it is effectively used for multiple time features along with several features. It takes minimum execution time by searching globally and also it updates new best similarity values quickly. Hence it increases the classification accuracy higher for the given specified datasets and best features are retrieved by using improved PSO algorithm more accurately.

2.7. Modified Artificial Bee Colony Algorithm (MABCA)

1. Initialize parameters as n, m, I, a and ec-length

n = Number of employed bees

m = Number of onlooker bees (m>n)

Iteration I : Maximum iteration number

a: initial value of penalty parameter for jth agent

ec-Length: Length of ejection chain neighborhood

2. Build primary employed bee colony results

For each bee evaluate fitness function f = $\sum_{j=1}^{m}\sum_{i=1}^{n} c_{ij}x_{ij} + \alpha \sum_{j=1}^{m} max\left\{\sum_{i=1}^{n} b_{ij}x_{ij} - a_j\right\}$

3. I =0

4. Repeat

5. N = 0

6. Repeat the process which is given below

7. If fit(ShiftNeighbour) <fit(EmployedBee) then

8. Employed Bee = Shift Neighbour

9. If fit(DoubleShiftNeighbour) <fit(EmployedBee) then

10. Employed Bee = DoubleShift Neighbour

ISSN (online) : 2277-5390
ISSN (Print): 2277-5382

11. Discover the probabilities using objective function Pi=$\frac{\Sigma(1/fit)^{-1}}{fit}$

12. Allocate onlooker bees to employed bees

13. For all Onlooker Bees

14. Ejection -Chain Neighborhood

15. Find best Onlooker, replace with respective Employed Bee

16. iffit(Best Onlooker)< fit(Employed)

17. Find best Feasible Onlooker, replace with Best solution,

18. if fit (BestFeas Onlooker)

19. N=N+1;

20. until (N=employed bee)

21. I = I+1

22. until (I=maxIteration)

The algorithm describes the concept of universal gravitation into the consideration of the affection among employed bees and the onlooker bees. By allocating diverse values of the control parameter, the universal gravitation is concerned for the artificial bee colony algorithm [19] while there are different quantities of employed bees and the single onlooker bee. Consequently, the investigation capability is converted about on typical in this algorithm.

2.8. Artificial Neural Network Algorithm (ANN)

Input: N training samples,
Output: predicted class
For each sample do
Input[i] =sample
Foreachi in neural metwork do
Output [i] = module. ForwardPropagate(input[i])
Input[i+1]=output[i]
End
Predictedclass = criterion (output)
If training then
Check error attributes
Foreach [k-i] in neural network do
Output = module. backwardPropagate
Input[i+1]=output[i]
End
End
End

For the number of input training sample we have to perform the analysis. For each input sample the neuron network perform the output sample based on the prior knowledge. It is used to search the more accurate results along with hidden layer and this layer is used to map the similarity for corresponding input sample. Hence it avoids the error values also it progress the speed of process more efficiently.

3. Results and Discussion

In this section, the overall performance metrics are evaluated using present and proposed methodologies. The performance metrics are such as recall, accuracy and precision. The existing random forest, support vector machine and improved PSO algorithm is used to classify the multiple measurement of specified dataset. However the existing

ISSN (online) : 2277-5390
ISSN (Print): 2277-5382

system has shown the lower performance in the classification results. The proposed Modified Artificial Bee Colony Algorithm (MABCA) has shown the higher performance in the classification results. The proposed MABCA with Transductive SVM (TSVM) and artificial neural network (ANN) [20] provides superior classification accuracy results. From the experimental result, we consider the performance and conclude the proposed system is better than the existing system. Thus the experimental results prove that the proposed method has high performance when compared with existing methods.

3.1. Accuracy

The accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

Accuracy can be calculated from formula given as follows

$$Accuracy = \frac{\text{True positive } + \text{ True negative}}{\text{True positive } + \text{ True negative} + \text{False positive} + \text{ False negative}}$$

An accuracy of 100% means that the measured values are exactly the same as the given values.

From the Figure 1, we can observe that the comparison of existing and proposed system in terms of accuracy metric. In x axis we plot the types and in y axis we plot the total accuracy values. The total accuracy values are lower by using existing algorithm of MMSVM and IPSO algorithm. The accuracy value of MMSVM and IPSO is 84% and 86% respectively. The accuracy value is higher by using the proposed of MABCA with TSVM and ANN. The accuracy value of MABCA with TSVM is 88% and MABCA with ANN is 90%. From the result, we conclude that proposed system is superior in performance.

From the graph, the values are tabulated in the Table 1. It shows that the proposed system is shown higher accuracy values compare than previous algorithm. Thus the result concludes that the proposed system is used to provide efficient classification results.

Figure 1. Accuracy

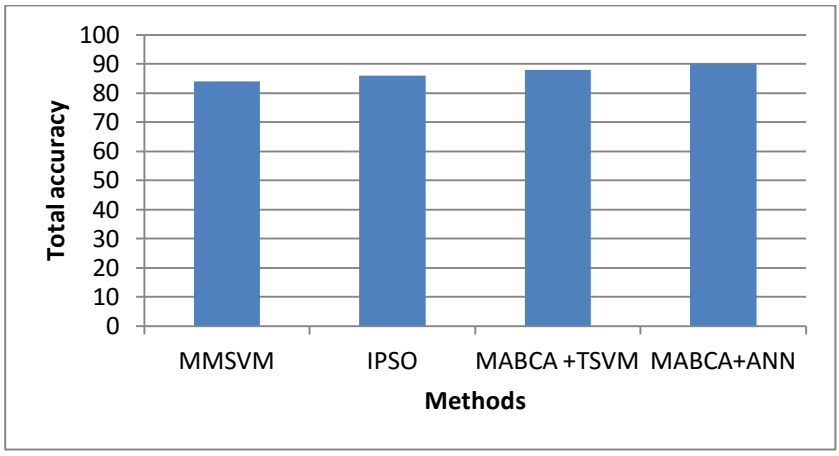

Table 1 .Accuracy

Performance metric	MMSVM	IPSO	MABCA+TSVM	MABCA+ANN
Accuracy	82%	85%	89%	90%

3.2. Precision

The precision is calculated as follows:

$$Precison = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}$$

ISSN (online) : 2277-5390
ISSN (Print): 2277-5382

Precision can be seen as a measure of exactness or quality, whereas recall is a measure of completeness or quantity. In simple terms, high precision means that an algorithm returned substantially more relevant results than irrelevant. In a classification task, the precision for a class is the number of true positives (i.e. the number of items correctly labelled as belonging to the positive class) divided by the total number of elements labelled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labelled as belonging to the class).

From the Figure 2, we can observe that the comparison of existing and proposed system in terms of recall metric. The Intermediate values can be calculate let f(x) be a continuous function on the interval (0.7,0.86).If d->[f(a),f(b)],then there is a c->[0.7,0.86] such that d=f(c) i.e ., f(c)=0.79.In x axis we plot the types and in y axis we plot the recall values. The recall values are lower by using existing algorithm of MMSVM and IPSO algorithm. The recall value is higher by using the proposed of MABCA with TSVM and ANN algorithm. From the result, we conclude that proposed system is superior in performance.

From the graph, the values are tabulated in the table 2. It shows that the proposed system is shown higher precision values compare than previous algorithm. Thus the result concludes that the proposed system is used to provide efficient classification results.

Figure 2. Precision

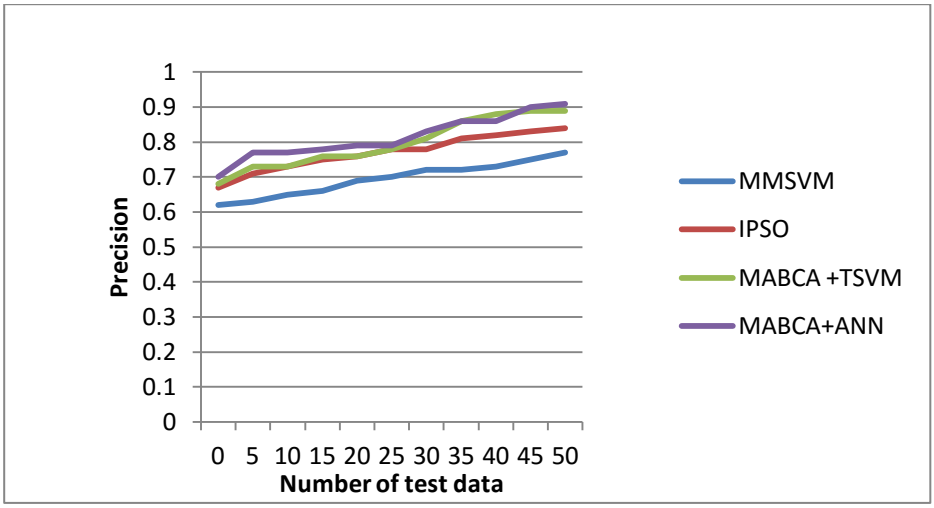

Table 2. Precision

Precision	MMSVM	IPSO	MABCA+TSVM	MABCA+ANN
0	0.62	0.67	0.68	0.7
5	0.63	0.71	0.73	0.77
10	0.65	0.73	0.73	0.77
15	0.66	0.75	0.76	0.78
20	0.69	0.76	0.76	0.79
25	0.7	0.78	0.78	0.79
30	0.72	0.78	0.81	0.83
35	0.72	0.81	0.86	0.86
40	0.73	0.82	0.88	0.86

4.3. Recall

The calculation of the recall value is done as follows:

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

The comparison graph is depicted as follows:

Recall is defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents, while precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search. Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e.

ISSN (online) : 2277-5390
ISSN (Print): 2277-5382

the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been).

From the Figure 3, we can observe that the comparison of existing and proposed system in terms of recall metric. The Intermediate values can be calculate let f(x) be a continuous function on the interval (0.64,0.77).If d->[f(a),f(b)],then there is a c->[0.64,0.77] such that d=f(c) i.e ., f(c)=0.68. In x axis we plot the types and in y axis we plot the recall values. The recall values are lower by using existing algorithm of MMSVM and IPSO algorithm. The recall value is higher by using the proposed of MABCA with TSVM and ANN algorithm. From the result, we conclude that proposed system is superior in performance.

From the graph, the values are tabulated in the table 3. It shows that the proposed system is shown higher recall values compare than previous algorithm. Thus the result concludes that the proposed system is used to provide efficient classification results.

Figure 3. Recall

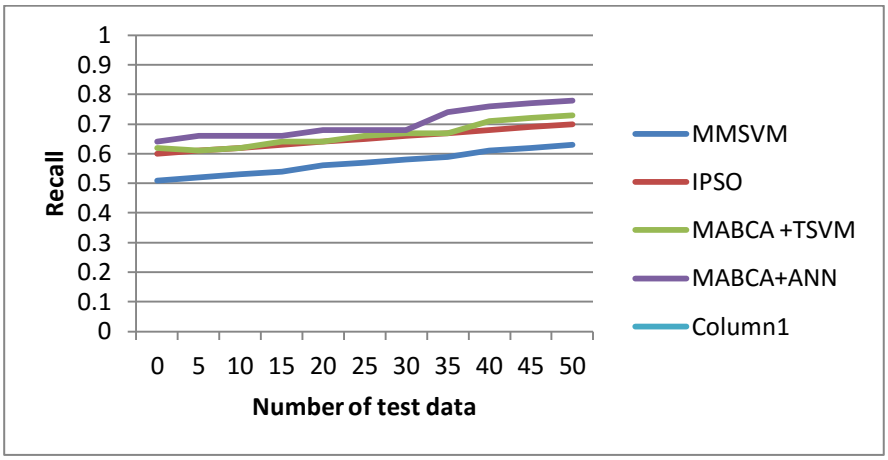

Table 3. Recall

Performance metric	MMSVM	IPSO	MABCA+TSVM	MABCA+ANN
Recall				
0	0.51	0.6	0.62	0.64
5	0.52	0.61	0.61	0.66
10	0.53	0.62	0.62	0.66
15	0.54	0.63	0.64	0.66
20	0.56	0.64	0.64	0.68
25	0.57	0.65	0.66	0.68
30	0.58	0.66	0.67	0.68
35	0.59	0.67	0.67	0.74
40	0.61	0.68	0.71	0.76
45	0.62	0.69	0.72	0.77

5. Conclusion

In this section, the conclusion decides that the proposed system is increased the classification performance using modified artificial bee colony algorithm. The various time series data is implemented and the methods are focused on the classification of more accurate results. The existing RF, SVM, MMSVM and IPSO algorithms are used to handle the multiple classification time series data and also dealt with the unbalanced dataset. The proposed MABCA with TSVM and ANN algorithm is used to improve the classification performance and reduce the time complexity issues significantly by using global optimal features. Thus, the experimental result proved that the proposed system is better than the existing system.

6. Acknowledgement

We the authors assure that, this is our own work and also assure that there is no conflict of interest.

7. References

1. N. Anwar, E. Huntz, W. Kolch, A. Pitti. Semantic Data Integration for Francisellatularensisnovicida Proteomic and Genomic Data. 2010; 21(5), pp.321-345.
2. I. Batal, H. Valizadegan, G. F. Cooper, M. Hauskrecht. A pattern mining approach for classifying multivariate temporal data.In Proceedings IEEE International Conference Bioinformatics and Biomedical. 2011; 20(3), pp. 358-365.
3. L. Breiman. Random forests, Mach. Learning.2001; 45(1), pp. 5-32.
4. Damian Bargiel, S. Herrmann. Multi-temporal land-cover classification of agricultural areas in two European regions with high resolution spotlight terraSAR-X data.2011; 3(5), pp. 859-877.
5. M. Campos, J. Palma, R. Marın, Temporal data mining with temporal constraints artificial intelligence in medicine. 2007; 2(4), pp. 67-76.
6. S. Dowdy, S. Wearden, D. Chilko. Statistics for Research', 3rd (edn). New York, NY, USA: Wiley. 2004; 42(6), pp. 625-640.
7. T. Exarchos, M. Tsipouras, C. Papaloukas, D. Fotiadis. An optimized sequential pattern matching methodology for sequence classification.2009; 19(2), pp. 249-264.
8. S. P. Gardner. Ontologies and Semantic Data Integration. Drug Discovery Today, 2005; 10(14), pp.1001-1007.
9. S. Gupta, D. Kumar, A. Sharma. Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis. 2011; 11(7), pp.125-135.
10. J. Han, M Kamber. Data Mining: Concepts and Techniques. 2nd ed. San Francisco, CA, USA: Morgan Kaufmann. 2006; 20(12), pp.325-365.
11. M. .A Hern´andez, S. J. Stolfo. Real-world data is dirty: Data cleansing and the merge/purge problem,' Data Mining Knowledge Discovery, 1998; 2(1), pp. 9-37.
12. Joseph Sexton, D. L. Urban, M. J. Donohue, C. Song. Long-term land cover dynamics by multi-temporal classification across the Landsat- 5 record. Remote Sensing Environment, 2005; 128(4), pp. 246-258.
13. C.H. Lee, J C Chen, Tseng V S. A novel data mining mechanism considering bio-signal and environmental data with applications on asthma monitoring. Computer Methods Programs Biomedicine, 2011; 101(1), pp. 44-61.
14. M. Lee, H. Lu, T. Ling, Y. Ko. Cleansing data for mining and warehousing,' in Database and Expert Systems Applications.1999; 46(6), pp.750-807.
15. M. Lenzerini, Data integration: A theoretical perspective, In Proceedings 21st Symposium on Principles of Database Systems, Madison, WI, USA, 2002; 35(7), pp. 233-246.
16. Y. Sam Sung Zhao, Li, Peng Sun. A Fast Filtering Schema for Large Database Cleansing, 2002; 26(8), pp.520-530.
17. J. Soni, D. Ansari, S. Sharma. Predictive Data Mining for Medical Diagnosis:An Overview of Heart Disease Prediction, 2011; 47(2), pp.640-650.
18. M. Stacey,C. McGregor. 'Temporal abstraction in intelligent clinical data analysis: A survey,' Artificial Intelligence in Medicine, 2007; 39(1), pp. 1-24.
19. R. Saranya, K. Balamurugan,M. Karuppasamy. Artificial Bee Colony Algorithm Based Congestion Management in Restructured Power System. *Indian Journal of Science and Technology*, 2015; 8(S7), 171-178.
20. A. Dhivya, S. N. Sivanandan. Hybrid Fuzzy Jordan Network for Robust and Efficient Intrusion Detection System. *Indian Journal of Science and Technology*, 2015; 8(34), pp. 1-10.

The Publication fee is defrayed by Indian Society for Education and Environment (iSee). www.iseeadyar.org

Citation:

Priyanga M, Dr. K. Sasi Kala Rani, Pavithra M, Yamunadevi S. The Multiple time series clinical data processing with modified artificial bee colony algorithm and artificial neural network. *Indian Journal of Innovations and Developments.* 2016; 5 (5), May.