INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 16, Issue: 6, Month: December, Year: 2021
Article Number: 4305, https://doi.org/10.15837/ijccc.2021.6.4305

CCC Publications

AGORA UNIVERSITY PRESS

Threshold based Support Vector Machine Learning Algorithm for Sequential Patterns

S. Imavathy, M. Chinnadurai

S. Imavathy
Research Scholar
Anna University,Chennai, Tamil Nadu, India
imavathy.sphd@gmail.com
coressponding author

M. Chinnadurai
Professor, CSE Department
E.G.S Pillay Engineering College, Nagapattinam, Tamil Nadu, India
mchinna81@gmail.com

Abstract

Now a days the pattern recognition is the major challenge in the field of data mining. The researchers focus on using data mining for wide variety of applications like market basket analysis, advertisement, and medical field etc., Here the transcriptional database is used for all the conventional algorithms, which is based on daily usage of object and/or performance of patients. Here the proposed research work uses sequential pattern mining approach using classification technique of Threshold based Support Vector Machine learning (T-SVM) algorithm. The pattern mining is to give the variable according to the user's interest by statistical model. Here this proposed research work is used to analysis the gene sequence datasets. Further, the T-SVM technique is used to classify the dataset based on sequential pattern mining approach. Especially, the threshold-based model is used for predicting the upcoming state of interest by sequential patterns. Because this makes deeper understanding about sequential input data and classify the result by providing threshold values. Therefore, the proposed method is efficient than the conventional method by getting the value of achievable classification accuracy, precision, False Positive rate, True Positive rate and it also reduces operating time. This proposed model is performed in MATLAB in the adaptation of 2018a.

Keywords: Data mining; Sequence Patterns; Threshold based Support Vector Machine Learning Algorithm; Classification Accuracy; Sequential mining;.

1 Introduction

The bioinformatics analysis is widely used in the area of computer science and applications, which deals with the data for collecting, organizing, and analysing. [1] Here the DNA and protein sequence are analyzed. Pattern mining algorithm with its sequential database are analysed with various domains. In this frequent data, a small gap in the pattern is restricted for finding valid pattern of dataset. The pattern recognition process involves the gap constraints, length constraints, maximum supports, and minimum type. Here the maximum support is no greater than the sub-support matrix. Youxi Wu and others uses non-overlapping sequence to specify the gap constraints by pattern mining approach. Here the gap constraints is obtained by finding transcriptional site of gene from DNA cells. Bioinformatics is also utilizes the area of Genomic signal processing for dealing with the digital signal processing (DSP) applications. [2] Clustering method is to identify the sequence of gene data, which is performed with online collaborative model. [3] Particle Swarm optimization technique with adaboost model had classified the result of DNA gene sequence by data mining approach. Identifying the species name or organism type is by gene sequence and classification techniques. This will provides the DNA-attributes, nature of DNA cell and the type of species of DNA. The data mining is used to extract the data from required field; in this, the subtopic performs the pattern mining approach for analysing patterns by statistical mode. Classification technique in data mining is mostly prefers machine learning algorithm. The sequential pattern analysis will discover the pattern of transcriptional gene data. This technique is to find the matches of frequently occurrence of data. To identify the pattern various measures are investigated. The minimum support matrix is to provide the threshold value to the sequence. The DNA sequence having the responsibility to regulate the gene, the characters A, G, T, and C is ordered with biological term s. the Adenine and thymine is paired with X mentioned and Cytosine and Guanine is paired with Y mentioned. In previous research work, [4] has described the dataset by performing the artificial neural network and deep learning algorithm. [5] Hadoop and weka distribution has proposed with machine learning algorithm and data mining techniques. Support Vector Machine (SVM) learning algorithm describes the classification of genes via intron and exon combination in the DNA cells. This classification process is based on size of given dataset. The boundary function of SVM is based on the value of threshold used for dataset, which determines the weight of classes of the boundary. After performing train function, the class of train and performance of test is validated. By determining kernel function, we can evaluate the class region of data. Based on the different threshold value assignment, the classifier will provide the various tested sequence of data. It is used in the linear classification method for analysing the larger datasets. After determining training nodes of sequence, the boundary will terminates to sort the value for arranging the nodes. Based on the database the classification is done with umber of sequences and list of identity, which evaluates the sequential dataset classifications. Here the sequence of data is identified as Is and dataset is represented as X_d. With the set of support matrix, we can determine the total number of sequence presented in the dataset.

$$\sigma(s) = |Is\epsilon Xd|s < Is| - - - - - - 1$$

With the ratio of 's', the number of sequence in the database is given by,

$$\sigma(s) = \frac{(|Is\epsilon Xd|s<Is|)}{Xd} - - - - - - 2$$

Here equation (2) represents the exact and relative support matrix is changeable by's' ratio. If the pattern sequence's' is greater or equal to the threshold value of dataset 'Xd', that expresses $\sigma(s) >= SminTHD$. Various node values are represented by kernel function 'k'. if the K > 0, the processor initialize the distribution factor. In pattern mining process, the k is always increments to the value '1' that is k+1. Sequence of data contains various items and it is performed with various classification technique that is GSP, SPADE, Prefix span, and GSpan. These techniques are used feature classification of SVM model. Minimum support matrix will fix the threshold value for mining the sequence patterns. The sequential data mining with classification is done by extracted dataset. Initially the dataset is analysed for its application and performs the sequential operation. Sampling process is done by getting the threshold value of each sequence. This high dimensional data is send to the training process and performs the operation of minimum support. It will generate the minimum valued data to the testing process and it minimizes the dimension of original dataset. After that by performing classification process, we can classify the dataset according to the genes sequence item sets. Transcriptional dataset is used for gene sequence determinations and it classified based on the mined data. Machine Learning technique is used for classifying the dataset by assigning class value to the train and test sets. Class boundary is selected in training set, it maximizes the support matrix's margin, and it will denote the length and distance of class boundary for getting compact data. DNA sequence has classified by integrating DM and ANN approach [4]. Auroral image sequences dataset is to presented and it is classified using pattern-mining techniques [20]. Nowadays, the machine-learning (ML) algorithm is a most widely used technique for data mining algorithms. Fingerprint based DNA sequence is to detect the present gene, identify the forensics, and testing the parental genes. Bioinformatics sequence used in machine learning algorithm provides better result as compared with other existing methods. Genome sequence having high volume of gene data, which is analyzed by monitoring number of data sequences. Transcriptional data sequence is used for classifying the gene function by specifying species.

This proposed research work is summarized as follows. Section II is the survey analysis of recent related literatures. Section III is the Existing method description and techniques used. Section IV is the proposed methodology-using threshold based support vector machine learning model for classification and gene expression. Section V is the result and discussion part. Section VI concludes the proposed method and provides the thought about future work.

2 Literature Survey

Gyula Dorgo., et. al., (2018) has proposed the sequential mining approach for alarm suppression applications. Statistical data is processed to find the abnormal condition of alarm. Here the suppression sequence is to set systematic analysis for processing and controlling the action. The technique uses bayes classifier with multi-temporal sequence mining algorithm. This method reduced the losses and suppression speed, since the operator will keep the operation in normal condition by indicating the present state display. Target and action limit is set with alarm rate with period, which is used to analysis the sequential data. Threshold value is assigned with statistical configurations of dataset. After performing the sampling process, the data sequence is initiated with threshold value for minimizing the dimension of data length. Alarm reduction model comprises fault detection and observation unit, which provides the result of work reduction and failure prediction. Here the data mining technique is used to extract the pattern of alarm suppression ratio. The alarm will triggers the next relay and it returns to next end state of sequence pattern. By the use of suppression in the sequential pattern mining, the alarm management and control strategy is performed and it reduces the data loss but increases the execution time.

Bao huynh., et. al., (2017) has proposed the parallel method is used to mining the sequential dataset with dynamic load balancing and dynamic bit vector options. The frequent data is performed in the closed sequential patterns and it is analysed by prefix span, which reduces the execution time of classifier. If the database contains the large set, the sequence will shows the exponential numbering of dataset and it is compact by data extraction process in data mining approach. Here the parallel method is used to solve the problem of expensive computations and large sequence analyser. Since it improves the processing speed but the number of stages increased while performing sorting process. Parallel architecture is modelled for performing sequential pattern analysis with dynamic bit vector and load balancing. Here the CPU time is changed based on the different architecture in parallel analysis. This method is design with various datasets and performed in Intel core i5-6200U with 2.3 Hz frequency and 3MB memory.

Po-Ming Law., et., al., (2018) has present the pattern mining technique for temporal query and sequence exploration of recursive operations. This method is not sufficient for support matrix analysis. While performing the event analysis, the sequence are extracted by providing the mined data in the output. By enabling the event sequence, have to identify the item sets. This makes the reduction in data loss on large sequence analysis. Sequential mining approach is to processed and controls the data for reducing the data loss. Since the classified sequences and segments are visualized in the database. Here the database is the marketing and health information analysis. By setting the time and by set the number of events in the constraints makes system effectiveness. The analysis process includes splitting and selecting the segments based on patterns, which depends on temporal data.

Seema Sharma., et., al., (2018) has present the machine learning technique for processing data by data mining technique. This survey analysis the knowledge based acquisition with its classification. Based on the assignment of class and group of given data, the predefined values are set for each instance and performs classification. Here the technique used is SVM, decision tree, and K-nearest neighbour with Bayesian network. By using decision tree for classification methods, the splitting strategy is to generate the gain value and it ratio. For finding the value of gain ration, we have to calculate the entropy and information gain. Here SVM technique uses K-nearest neighbour node, which reduces the data loss while performing classification. With various techniques and determinations, the computational cost is increased for large sample sets, which is very sensitive.

Anuja Jain., et., al., (2017) has proposed the data mining technique for hadoop with weka distributions. Here the supervised Machine Learning (ML) algorithm is used for classification. For big data mining technique the Hadoop approach is used and it is connected with Weka tool. ML algorithm uses Naive Bayes and SVM for big data analysis and it compared the result of accuracy for raw data versus normalized data. By this work, it handles the problem of big data analyses, which are data integration, data volume, technical skills requirement, and cost solution. Here the Hadoop uses Apache open sources, which allows the system for distributed processing model. By clustering process specification of characteristics is assigned by predefined classes. Transactional data is analysed by sequential data processing technique, which is created by decision tree structure. Reinforcement learning technique is to connect with dynamic setting data, which provides the feedback to the user for rewards and punishments.

Sadok Rezig., et., al., (2018) has proposed the technique for predicting maintenance activity by sequential data sets. The data-mining algorithm used for various applications, which bases the spare parts maintenance. Threshold value is assign for support matrix is set previously and by this information, we can finds the sequential data patterns. Frequent data mining is combined with process of classification, prediction, and clustering. Threshold support value provides the pattern assignment of each data to be processed. By this technique, we can analyse the raw data for removing the unwanted data this will reduce the data loss. Sequential pattern is used to generate the raw data into sequential model, which is known as antecedent and consequence. Here they takes the spare part maintenance of a company, which provides the maintenance parameter, spare part code and description about the spare part. By this parameters they can obtain the sequence, intervals by weeks and percentage of supports. Final determination is about classification with its related parameters.

Cheng Zhou., et., al., (2016) has presents the sequential classification method, which is based on data mining with pattern recognition techniques. Here the feature vector based model is used to generate the sequence for classification. The efficient technique is to composed pattern mining technique takes the interesting data to label the class value, which is based on support matrix and classification rule. Some situations, the dataset is not allocated properly in an order, this leads to data loss. However, by the use of sequence classification rule, we can remove the unwanted data. Since it, obtained the mined data and accuracy of classification techniques. Naives Bayes classification technique is used in pattern mining approach for processing sequential dataset. Interesting dataset pattern is based on support and cohesion. By comparing with predictive accuracy, it will not support for big data mining approach.

Sathish Kumar S., and Duraipandian N., (2012) has presented the DM and ANN based classification technique for DNA sequence analysis. To identify the species from DNA gene sequence, this is considered as a raw dataset. If the dataset contains unwanted information, it is neglected for classification process. Here the artificial neural network model is used for data mining approach. High dimensional dataset is mined by principal component analysis strategy and it is validated by 10-fold cross validation approach. By this statistical validation, the classification model is performed. By this bioinformatics data, they can obtain the species name by DNA cells and classification accuracy. Genomic signal processing tool is used to classify the dataset. This species classification model is depends on nucleotides and data is mined by nucleotide pattern. By calculating various pattern lengths, can obtain the parameters assignment as A, T, G, and C. In training phase, multi dimensional structure is reduced by using artificial neural network. In support matrix, species brucella suis and C-elegans are determined. Classification technique improved the accuracy rate.

3 Existing Method

Plenty of research ideas are focusing with the performance of various mining techniques for bioinformatics datasets. Data mining technique is the most emerging technology, which is used for various applications. The conventional method used the artificial neural network (ANN) for classification. DNA splice junction sequence is used for data analysis. While performing sequential pattern mining, initial stage having larger data dimension, this should be reduced as small data dimensions. By comparing with artificial neural network classifications, the proposed research work improves the accuracy and truth factor. Gene sequence is associated with genome for assigning the genetic information about the species. Most of the conventional method uses ANN, genetic algorithm, and fuzzy logic system with the structure of hybrid model

3.1 Neural Network based Pattern mining Technique

Neural network classification method is used in the previous research work [1]. With data mining technique, the sequential pattern recognition had performed. The Radix Tree miner algorithm is used to structuring the model. Here the sequential pattern mining technique is employed by performing larger dataset. The neural network model is performed to classify the dataset. In this data set, A, C, G, and T are the nucleotides, which assigning the value of each item sets for performing classification in NN. By Radix tree miner, if the sub-trees are not divided for upcoming process have to neglect the set because it is an unwanted data. This takes number of stages high than the proposed T-SVM based pattern miner technique. Execution time of classification technique is increased by increasing number of hidden layers in the NN. When compared to the previous work with proposed T-SVM based pattern mining technique, the execution time is reduced in proposed method, this makes the system efficiency. Figure 1 shows the block diagram for existing Neural Network classification for sequential pattern mining for Splice gene sequence dataset. Initially the raw data set of DNA cells are collected from various sources. It has high dimensionality in nature. This is reduced by various techniques like radix tree miner algorithm and artificial neural network for classification. The algorithm had evaluate the sequence for identifying pattern. In support block, the threshold value of each sequence value allocated for minimizing the sequential pattern dimensions. For each data set, the sequence ID is given with respect to the pattern alignment. The length of sequences is increased by time complexity of radix tree miner algorithm. After performing radix tree miner the high dimension data set is sent to the training process. Prefix span method is also used for making pattern of gene sequence, but

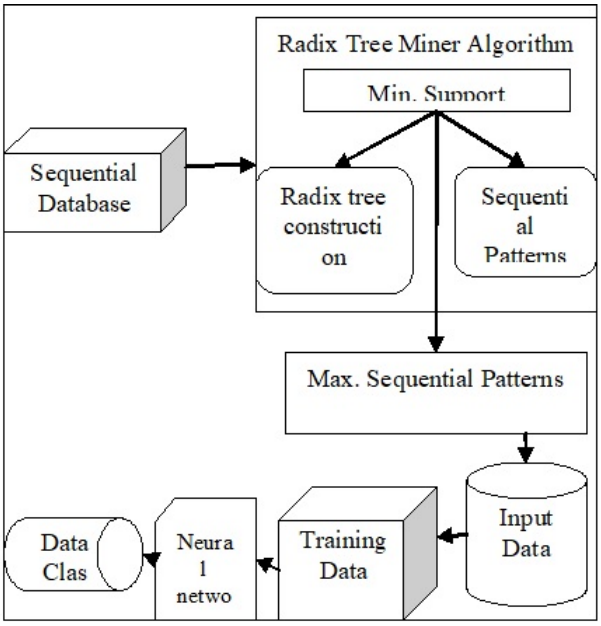

Figure 1: Block Diagram for NN for sequential pattern mining technique

it consumes much timing to complete the process and requires more memory to process from the projected data sequence.

4 Proposed T-SVM Based Sequential Pattern Mining Techniques

The proposed threshold based Support Vector Machine learning algorithm is preferred in this research work for classification of sequential pattern mining approach. The proposed method takes the data set of splice-junction gene sequence as mentioned in previous research work data set. Here the DNS sequence contains the four variable set of nucleotides, that is Adenine 'A', Guanine 'G', Cytosine 'C', and Thymine 'T'. These species are formed as nucleotide. Initially the long sequence set to be processed for reducing the dimension of dataset, this makes the system efficiency. The proposed method is integrating the data mining technique uses Machine Learning (ML) algorithm for classification. To generate the length of each sequence, we can obtain the value of threshold. By the statistical data, we can obtain the classification in linear based supervised machine learning algorithm. Figure 2 shows the block diagram of proposed model with T-SVM algorithm and sequential pattern mining techniques. First, the dataset of Splice junction gene sequence is taken for this analysis. In minimum support block, the threshold value is set for initial data sequence, which is based on rule based mining technique. The gene sequences are mined by pattern mining techniques. Based on the threshold value, have to allocate the pattern for training process. After performing training process, the SVM model is to generate the class label to the sequential data for performing the testing process. Finally, the test sequence is allocated for performing classification based on its threshold value. However we can obtain the sequentially pattern mined data.

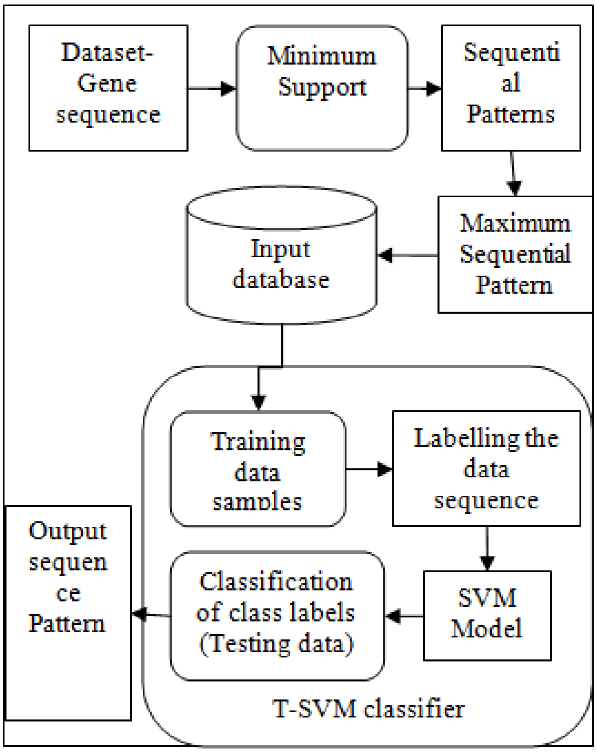

Figure 2: Block Diagram of proposed T-SVM for sequential pattern miner system

A.Dataset

Pattern mining process is performed for data set of Splice junction gene sequence of primates via DNA cells, which is taken from centre of machine learning and intelligent systems and it is associated with imperfect domain theory. The splice sequences are pointed from DNA sequences and it recognizing with two boundaries, called intron and exon. Here 3190 instance is capture with 61 attributes. The below table 1 mentioning the nucleotides arrangements in datasets indicates the variables D, N, S, and R. these raw dataset is generated for performing sequential pattern mining technique. This dataset

Table 1: Dataset character mentioned

Assigned CHARACTER	NUCLEOTIDES Order
D	A G T
N	A G C T
S	C G
R	A G

is capable of 10-fold validation techniques, which is applied to the 60 sequential DNA nucleotide positions. Class distribution is mentioned as EI and IE of intron/exon. Genetic information about the DNA sequence is given as ,

1. Molecular biology
2. Evolutionary biology
3. Metagenomics

This information is based on genome and proteins in the DNA sequence. This nucleotides collection of gene is the primates splice junction sequence, which is identify by potential targets, study with differ organism in evolutionary biology, and species representation in metagenomics. The transcriptional dataset is analyzed by three different analysis that is microarray analysis, tiling arrays, and regulatory sequence analysis.

B.Sequential Pattern mining Technique

Machine learning algorithm uses pattern-mining techniques with sequential data. Hidden parts of sequences should be mined by ML algorithm. This technique will control and detect the pattern for classification. Based on the sequential data, the statistical analysis is taken for classification model and it is extracted for respective applications. These sequential patterns are identified by assigning threshold value of each sequence of data and it is from transactional database. Dataset is frequently analyses the pattern for pattern arrangement and it fills sufficient to the algorithm. This is for predicting the respective pattern recognition and classification models. Based on pattern mining techniques, the pattern is allocated by threshold value and it minimizes the dimension of pattern. It is completely depends on outcome of the model. The distance of sequence in SVM is calculated by,

$$Distance(i) = (2 * m - 2 * count)/(2 * m - count) - - - - - - - - 3$$

Where, I is the iteration and 'm' is the length of input sequence. By pattern mining is a function is expressed as sigma that is given by

$$F_w, s(X) = sgn(wTX + s) - - - - - - - - 4$$

$$\mathrm{F}_{\mathrm{support}}\ (\mathrm{X}) = \mathrm{sgn}\ (\textstyle\sum_{i-1}^{n} b\ \lambda i\ Xi + s)\text{--------}5$$

Here this pattern is recognized from given input data sequence. Sequential pattern is identified as S, which s presented by threshold value S= S1, S2,. . . . . . ,Snand dataset is the sequential arrangement of nucleotides that is A, C, T, and G is given by, Here S1 and S is represented as,

$$\mathrm{S} = \frac{Xd}{a} S1 + \frac{xp}{a} S2 \text{ ---------}6$$

$$S1 = 1/X_d \sum(a_t - \mu 1)(a - \mu 1)^T - - - - - - - - - - 7$$

$$S2 = 1/X_d \sum(a_t - \mu 1)(a - \mu 1)^T - - - - - - - - - - 8$$

For k=1 the confusion matrix is,

$$\begin{bmatrix} 569.4 & 63.4 & 58.2 \\ 116.9 & 521.7 & 53.4 \\ 424 & 326.8 & 739.2 \end{bmatrix}$$

Confusion matrix is in the order of EI, IE, and Numbers of training region and the value is based on class test and gene train function. This confusion matrix provides the better accuracy result to the classification system. Based on the threshold variations in the input sequence, the arrangement of pattern is by labeling each sequence. This variation leads to changes in resultant pattern of every instance. This class of information of each sequence is the mined pattern. Based on the data sequences, the pattern-mined data is shown in Table 2. By the confusion matrix changes, the varying threshold value to each sequence is mined in a ordered pattern. Highlighted sequences are characterized as D, N, S, and R of nucleotides and it is mentioned in table 1. Probability of intron and exon is calculated by,

Table 2: Sequential Pattern mining techniques

K	**SEQUENTIAL PATTERN**
1	’AAGCCCATCCTAGAGAAGCTGACCCAGGACCAGGATGT GGACGTCAAATACTTTGCCCAG’
2	’ACGGAGCGAGTCTGGAACCTGATCAGATACATCTATAACC AAGAGGAGTACGCGCGCTAC’
3	’TGCTCTCCCAGGTCTACCCTGAACTGCAGATCACCAATGT GGTAGAAGCCAACCAACCAG’
4	‘TGCCTCCTTTCACACTCCTCTTGGGGCTCGTGACATTACG AACCCTAACCCGGGCCCTGC’
5	’TCGTGGCGTTTGTGGCAACCCCGGACACGGGGCACCAGC CAGTCAGCGGAGCCTCCTCAC’
6	’CATCGTCTACCTGGGTCGCTCAAGGCTTAACTCCAACACGC AAGGGGAGATGAAGTTTGA’
7	’GCCGCTTCCTCATCCTGGCACACTCTCTTCACAGCCGAAGA AGGCCAGTTGTATGGACCG’
8	’CGCACCTGGGCGCCCTGCTGGCAAGATACATCCAGCAGGC CCGGAAAGGTAAGAATGCTG’
9	’GGCCAGATCGTGCCATAGCACTCCACTTTGGGTGATAGAGG GAGACTCTGTCTCAAAAAA’
10	’GGAGTGGGGGCGGTGCGTCCTCCGGCCGGCAGCGGTGGC CACAGCTCTCCTCCCGCCGCC’

P_EI=numbers of EI /length of class

P_IE=numbers of IE /length of class

Based on the above equation model, we can obtain the probability value of EI and IE. By length of class and number of intron and exon will show the gene class of train function. The probability mining is given by,
For Sum the probability of ‘a’ is given by,

$$P(a) = \sum_b p(a,b) - - - - - - - - - - - - 9$$

For multiplication of the probability is given by,

P(a,b) = p(b|a) p(a)——10

The Bayes theorem is used to represents the pattern recognition in T-SVM. It is given by,

$$P(b/a) = \frac{(p(a/b)p(b))}{(p(a))} ------11$$

The gene sequence of primates is having attributes, nature and species types, which is classified based on the threshold based support vector machine-learning algorithm. Initially the nucleotides sequence pattern is mined by pattern mining, but it have a high dimensionality in nature. It is reduced by classifying the sequence of each sub sections. After performing training, the sequence should be allocated for testing; this provides the classified result by labeling the each class values.

C. Threshold based SVM Model

Proposed threshold based SVM model is used for classification of sequential patterns. Based on the statistical analysis of proposed method, we can classify the result. Kernel based classification model is structure for train and test sequence. The train data is linearly separable by pair of weight function and support. That is function of <w,s> is given by,

$$f(x) = f(W^T X + \lambda_0) ------- 12$$

$$W^T X + s >= 1, for X \varepsilon pi(pos)$$

$$W^T X + s >= -1, for X \varepsilon pi(neg)$$

Here the weight function 'w' is represented as,

$$W_p = \Sigma^{-1} \mu_p \text{ -------- } 13$$

$$W_{p,0} = -1/2\ \mu_p \Sigma^{-1} \mu_p + P(a) \text{ --------- } 14$$

The above represented functions are mined with DNA cells of primates. This linear function generalized with weight vector and data miner variable. Here the combination of nucleotides is mined for sampling the sequence, which have threshold value of each class of sequence. The labeling of each sequence is assumed for class value of the DNA. By extracting this sequence at each instance, the DNA gene formation is made for the species recognition. Here the support matrix is used to assign the threshold value of given sequence. After it is given to the training process of SVM model, the kernel function in T-SVM classification result is improved. The gauss distribution function is represented by single mean valued sequence that is given by,

$$G(a|\mu,\sigma^2) = \frac{1}{(\sqrt{2\pi\sigma 2})} \exp\{-1/2\sigma(a-\mu)^2\} \text{-------}15$$

The probability distribution function of Gaussian distribution model is given by,

$$P(i_a) = \varsigma P(i_a, i_b) di_b --------16$$

This proposed Threshold-SVM model is used to assume the threshold value at internal sub-section of sequence having threshold value by sampling process. After getting the extracted data sequence, the pattern is set by pattern mining technique for describing DNA attributes. This will determines the support value for constructing confusion matrix at each value of 'k'. After completing the process of training and testing, the mined dataset is classified. Table 3 represents the sequential pattern of training and testing process with its execution time. Here the gene sequences are aligned with respective threshold values for classification.

5 Results and Discussion

Thus, the sequential pattern mining technique was applied in primate splice junction gene sequence dataset and it is classified using threshold based support vector machine learning algorithm. Based on the dataset parameters, the classification process is performed. This achieves the result of classification

Table 3: Train and Test gene sequence alignment

TRAIN GENE SEQUENCE	TEST GENE SEQUENCE
'ATTCAAACAGCGCCTCAGACTACTTCATTTGG TACAAACAAGAATCTGGAAAAGGTCCTC'	'TGGACCATCGCGGATAGACAAGAACCGA GGGGCCTCTGCGCCCTGGGCCCAGCTC TGTCC'
Elapsed time is 1.308294 seconds.	Elapsed time is 5.475533 seconds.

accuracy and reduces the execution time. Table 4 represents the determination value of proposed T-SVM model using SPM technique.

Performance metrics

The evaluation of whole system is work with various aspects like accuracy, sensitivity, and precision.

a.Classification Accuracy

By finding the confusion matrix, we can calculate the accuracy by correctness value of proposed systems and it is associated with the value of true positive, true negative, false positive, and false negative in matrix form. Here the true positive is the rate of actual value to the predicted value of data. True negative is the actual and predicted rate is with class zero valued function. False positive is the rate of false rate of actual value versus predicted value. False negative is the actual class value versus predicted false value zero. The classification accuracy is calculated as,

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

b.Sensitivity

In pattern recognition, the sensitivity of mined pattern is depends on actual sequence and trained sequence with its threshold value. The sensitivity is also known as recall, which is calculated by actual value to the false negative value.

$$Sensitivity = \frac{TP}{(TP+FN)}$$

c.Precision

Precision value is about the capturing the correctness of the given sequence. The probability of getting the value of actual value into predicted value versus test sequence. It is formulated as given equations,

$$Precision = \frac{TP}{(TP+FP)}$$

Table 4: Values obtained in proposed T-SVM using SPM technique

Parameters	T-SVM using SPM method
Accuracy	90.64%
Sensitivity	100%
Precision	1
Execution Time	677ns (Train and Test)
Number of Pattern mined	768
Correct Sequence	1879

Figure 3 represents the comparison result of proposed method's number of patterns mined for an instance with conventional methods. Here it achieves the best result as compared with previous techniques. For every instances, the variation in threshold value changes the sequence, here this above result is based on the initial k=1, the sequence are "AAGCCCATCCTAGAGAAGCTGAC-CCAGGACCAGGATGTGGACGTCAAATACTTTGCCCAG' is mined. Here the function Gauss and kernel is represented for training and testing process. Figure 4 shows the graph model of Gauss and Kernel representations of T-SVM model. The below figure 5 is the comparison result value of classification accuracy, compared with conventional radix tree miner and Prefix span method. Figure 6 is the comparison result representation of compilation time with the unit of ns. The whole comparison result is shown table 5, which contains the classification accuracy, time taken to complete the

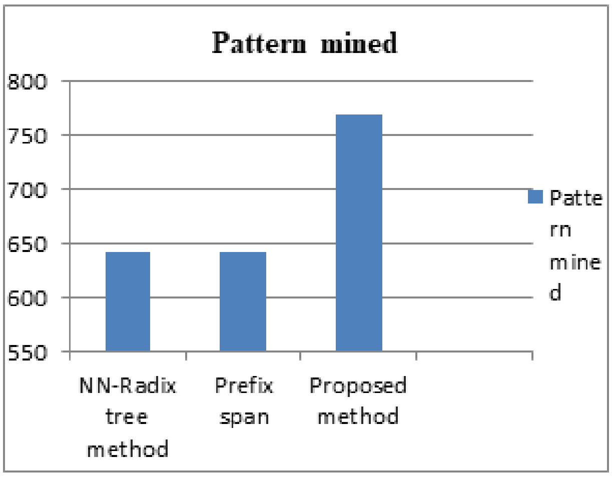

Figure 3: Comparison with the result of Number of pattern mined of an instance

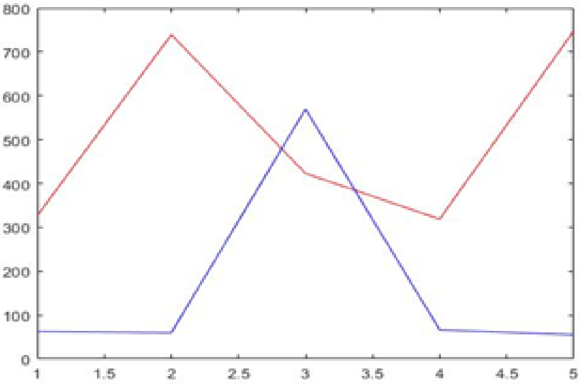

Figure 4: Gauss (red) and Kernel (blue) function

process, and number of pattern mined for an instance., which is compared with existing radix tree miner and prefix span methods. The proposed method achieves the better result as compared to the previous work. The above figure 7 shows the comparison results graph model; by this values, the proposed method achieves the result in the way of improved accuracy, reduced execution time, and number of pattern mined at an instance. Since the proposed method is efficient than the conventional method.

6 Conclusion and Future Scope

Thus, the proposed research work is concluded with the technique of sequential pattern-mining approach uses threshold based Support Vector Machine learning (T-SVM) algorithm for classification of DNA gene sequences. Here the pattern mining technique is used to model the given gene sequence by pattern mining method. Based on the data set attributes the gene sequence of DNA is mentioned. The high dimensional data set is reduced by performing training and testing of data. Based on the

Table 5: Comparison results

Parameter	**Neural network based Radix tree miner [1]**	**Prefix span based method**	**Proposed T-SVM method**
No. of Pattern mined	642	641	768
Accuracy	90.38%	74.83%	90.64%
Execution Time	795ns	907ns	677ns

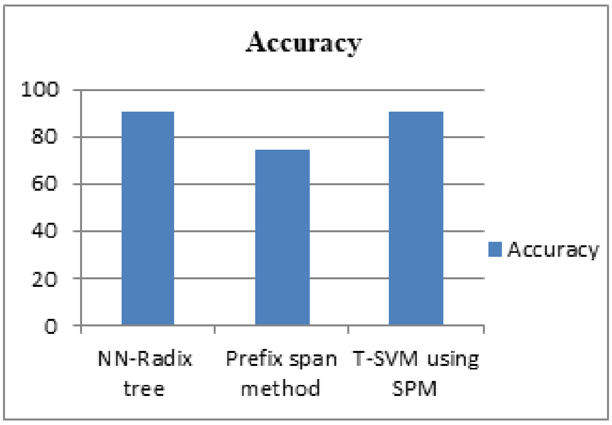

Figure 5: Comparison result of Classification Accuracy

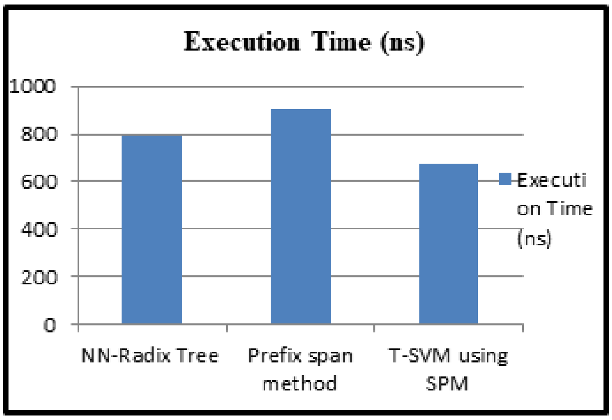

Figure 6: Comparison Result of Execution time of conventional methods versus proposed method

'k' value, the sequence of mined data is varied and it is given to the class database. By comparing with various classification techniques, the proposed method achieves the result by classification accuracy, number of pattern mined and execution time. In future work, the proposed method is enhanced by modifying novel techniques for various datasets with most achievable performance metrics. Classification accuracy is improved by providing correct sequence to the model.

Funding

No funding is involved in this work.

Conflict of interest

Conflict of Interest is not applicable in this work.

Authorship contributions

There is no authorship contribution.

Acknowledgement

There is no acknowledgement involved in this work.

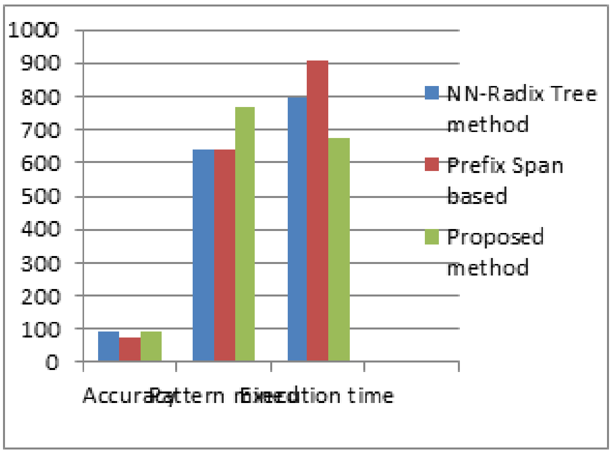

Figure 7: Comparison Result of Execution time of conventional methods versus proposed method

References

[1] K Poongodi., and A K Sheik Manzoor, (2019) .“Sequential Pattern Mining using RadixTreeMiner algorithm and Neural Network based Classification,” International Journal of Applied Mathematics and Information Sciences, vol. 13, No. S1, pp. 1-15,.

[2] Dilhan Perera., Judy Kay., Irena Koprinska., Kalina Yacef., and Osmar Zaiane.,(2009). “Clustering and Sequential Pattern Mining of Online Collaborative Learning Data,” IEEE Transactions on knowledge and Data Engineering, Vol. 21, No. 6, pp.759-772.

[3] Chieh-Yuan Tsai., and Chih-Jung Chen. (2015). “A PSO-AB classifier for solving sequence classification problems,” Applied Soft Computing, Vol. 27, pp. 11-27.

[3] Sathish Kumar S, and N.Duraipandian., (2012). “An Efficient Identification of Species from DNA Sequence: A Classification Technique by Integrating DM and ANN,” International Journal of Advanced Computer Science and Applications, Vol. 3, No. 8, pp. 104-114.

[4] Gyula Dorgo., and Janos Abonyi., (2018), “Sequence Mining based Alarm Suppression,” IEEE Access, Vol. 6, pp. 15365-15379.

[5] Bao Huynh., Bay Vo., and Vaclav Snasel., (2017) “An Efficient Parallel Method for Mining Frequent Closed Sequential Patterns,” IEEE Access, Vol. 5, pp. 17392-17402.

[6] Po-Ming Law., Zhincheng Liu., Sana Malik., and Rahul C. Basole., (2019) “MAQUI: Interweaving Queries and Pattern Mining for Recursive Event Sequence Exploration,” IEEE Transactions on Visualization and Computer GraphicsVol. 25, No. 1, pp. 396-406.

[7] Tiantian Xu., Tongxuan Li., and Xiangjun Dong., (2018) “.Efficient High Utility Negative Sequential Patterns Mining in Smart Campus,” IEEE Access, Vol. 6, pp. 23839-23847.

[8] Jinsong Zhang., Yinglin Wang., Chao Zhang., and Yongyong Shi., (2016) “Mining Contiguous sequential Generators in Biological Sequences,” IEEE Transacions on Computational Biology and Bioinformatics, Vol. 13, No. 5.

[9] Zhongliang Fu., Zongshun Tian., Yanqing Xu., and Kaichun Zhou., (2017) “Mining Frequent Route patterns Based on Personal Trajectory Abstraction,” IEEE Access, Vol. 5, pp. 11352-11363.

[10] Anuja Jain., Varsha Sharma., and Vivek Sharma., (2017) “Big Data mining using machine learning approaches for Hadoop with Weka distribution,” International Journal of Computational Intelligence Research, Vol. 13, No. 8, pp. 2095-2111.

[11] Seema Sharma., Jitendra Agrawal., Shikha Agarwal., and Sanjeev Sharma.(2013) “Machine Learning Techniques for Data Mining: A Survey,” IEEE International Conference on Computational Intelligence and Computing Research.

[12] Kapil Sharma., Ashok., and Harish Rohil.(2014). “A Study of Sequential Pattern Mining Techniques,” International Journal of Engineering and Management Research, Vol. 4, No. 1, pp. 241-248.

[13] Sadok Rezig., Zied Achour, and Nidhal Rezg.(2018) “Using Data Mining Methods for Predicting Sequential Maintenance Activities,” Applied Science, Vol. 8, pp. 1-13.

[14] S Rajasekaran., and L Arockiam., (2014) “Frequent contiguous Pattern Mining Algorithm for Biological Data Sequences,” Inter. Journal of Computer Applications, Vol. 95, No. 14, pp. 15-20.

[15] Sandeep R Suthar., Vipul K Dabhi., and Harshadkumar B Prajapati., (2017) “Machine Learning Techniques in Hadoop Environment: A Survey,” IEEE International Conference on Innovations in Power and Advanced Computing Technologies.

[16] Ya-Bo Liu, and Da-You Liu., “Mining Attributes Sequential Patterns for Error Identification in Data Set,” IEEE International Conf. on Machine learning and Cybernetics, pp. 1931-1936

[17] Cheng Zhou., Boris Cule., and Bart Goethals., (2016) “Pattern based Sequence Classification,” IEEE Transactionson Knowledge nd Data Engineering, Vol. 28, No. 5, pp. 1285-1298.

[18] Qiuju Yang., Jimin Liang., Zejun Hu., and Heng Zhao., (2012) “Auroral Sequence Representation and Classification using Hidden Markov Model,” IEEE Transactions on Geoscience and Remote Sensing, Vol. 50, No. 12, pp. 5049-5060.

[19] K. S. M. Tozammel Hossain., Debprakash Patnaik., Srivatsan Laxman., Prateek Jain., and Chris Bailey-Kellogg., (2013) .“Improved Multiple Sequence Alignments using Coupled Pattern Mining ,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 10, October 2013, No. 5, pp. 1098-1112,.

[20] Youxi Wu., Yao Tong., Xingquan Zhu., and Xindong Wu., (2018) “NOSEP: Nonoverlapping Sequence Pattern Mining with Gap Constraints,” IEEE Transactions on Cybernetics, Vol. 48, October 2018, No. 10, pp. 2809-2822.

[21] Ronald A Skoog., Thomas C Banwell., Joel W Gannett., Sarry F Habiby., and Marcus Pang., (2006)“Automatic Identification of Impairments using Support Vector Machine Pattern Classification on Eye Diagrams,” IEEE Photonics Technology Letters, Vol. 18, No. 22, pp. 2398-2400,.

[22] Siyabend Turgut., Mustafa Dagtekin., and Tolga Ensari., (2018), “Microarray breast cancer data classification using Machine Learning Methods,” IEEE Electric Electronicxs, Computer Science, Biomedical Engineering.

[23] Christopher M. Bishop., (2006) “Pattern Recognition and Machine Learning,” , pp. 1-758, Springer.

[24] Peter Wlodarczak., Jeffrey Soar., and Mustafa Ally., (2006)“Multimedia data mining using deep learning,” IEEE Inter. Conf. on Digital Information Processing and Communications.

[25] Jian Pie., Jiawei Han., B Mortazavi-Asl., Jianyong Wang., H Pinto., Qiming Chen., U Dayal., and Mei-Chun Hsu, (2004) “Mining Sequential Patterns by Pattern-growth: the PrefixSpan Approach, IEEE Transactions on Knowledge and Data Engineering, Vol. 16, Issue 11, pp. 1424-1440.

[26] Dhyanesh K. Parmar., Yagnik A. Rathod., and Mukesh M. Patel., (2014) “Survey on high utility oriented sequential pattern mining,” IEEE International conference on Computational Intelligence and Computing Research.

[27] D Martens., B B Baesens., T Van Gestel., (2009), "Decompositional Rule Extraction from Support Vector Machines by Active Learning," IEEE Transactions on Knowledge and Data Engineering, Vol. 21, Issue 2, pp. 178-191.

[28] Solomon H Ebenuwa., Mhd Saeed Sharif., Mamoun Alazab., Ameer AI-Nemrat., (2019). "Variance Ranking Attributes Selection Techniques for Binary Classification Probem in Imbalance Data," IEEE Access,24649-24666.

Copyright ©2021 by the authors. Licensee Agora University, Oradea, Romania.
This is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International License.
Journal's webpage: http://univagora.ro/jour/index.php/ijccc/

Member since 2012
JM08090

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

Cite this paper as:
Imavathy S.; Chinnadurai M. (2021). Threshold based Support Vector Machine Learning Algorithm for Sequential Patterns, *International Journal of Computers Communications & Control*, 16(6), 4305, 2021.
https://doi.org/10.15837/ijccc.2021.6.4305