An Enhanced Generalized Sequential Pattern Classification for Sequence Datasets

Immaculate Mercy A, Chidambaram M

Abstract: A Niche approach for classifying sequence Datasets is achieved by the use of EGSP (Enhanced Generalized sequential pattern) algorithm. The EGSP brings out a prediction model for working with the Sequence Datasets in the domain of Gene and Protein datasets. The Method proceeds by the way of generalizing the datasets of both the supervised and Semi-supervised data. The generalization brings out the candidate sequences which paves path for a distinct component extraction. The sequences are generated based on the threshold value which is then followed by applying the EGSP algorithm which brings out the sequential pattern from the pruned sequences. The extracted sequential pattern is then clustered using a gene clustering algorithm. The algorithm (MNBC) Modified Naïve Bayes Classification computes the probabilistic components for each class. The accuracy obtained is far better than the traditional classification algorithms. The resultant classification provides a solution for prediction methods for the selected domain and its applications. The algorithm used gives an upper hand over the computational costs which have been drastically minimized over the existing methods.

Keywords: EGSP, MNBC, Candidate sequences, Gene Clustering, Generalization.

I. INTRODUCTION

The motivation for the niche approach for classification of Gene datasets is twofold. The need for achieving a better classification leads to prediction of high dimension data which may be structured or unstructured in nature. Next the need for computation overcomes the deficiencies of the existing works. This twofold approach provides a way for better classification preserving the scalability and reliability factors.

Classification of Gene expression datasets paves way for channelizing the predictive analysis in the domains of Medicine, Bioinformatics, and Health Information systems. Various computational technologies for the Gene science and Bio-Sciences have evolved in the past decade which has provided numerous channels for stratifying data pertaining to a specific framework. The bottleneck in these approaches mainly lies in the fact of integrating data from different sources. Analysis of the data collected from different sources gives rise to complexity measures as the acquired data comes in various forms. The pre-processing of these data leads to a larger time complexity. However by the use of multi-variant methods these shortcomings have been surpassed. Thinking on the perspective of the dimensionality of the data the existing methods such as Support Vector Machines, KNN has proved to be efficient only on a selected smaller range of data.

In case of Supervised Machine learning algorithms a previously collected dataset together with the evidence is used to holster a decision model.

After it is holstered the characterization is done with the training dataset. As the datasets are large it is prone to curse of dimensionality which hinders the scalability and reliability of the devised method. The semi-supervised classification focuses on the analysis of the dataset and it is self contained by the use of KNN method. Though the KNN method tries to calculate the distance with the probability that a given instance belongs to a class it is almost based on the limited conditional probability.

Though all the existing techniques tried to prove on the factors of accuracy it has somewhere restricted on the fronts of scalability which poses a challenge on the reliability factor. In pursuance of ruling out the deficiencies in the existing works a new approach is devised which leads to a better efficiency on the Classification of High dimensional gene datasets. The framework that is proposed guarantees the fact that the EGSP provides better classification.

II. METHODS

There have been numerous classification techniques in the earlier studies but all those works has been restricted to a certain number of data sets and have proved the accuracy and efficiency on those datasets. Considering the nature of the data and the sources from which they have been obtained the earlier methods doesn’t suffice to the accuracy and the efficiency measures. The earlier machine learning algorithms have proved to work better only for a limited dataset. The data that has been considered for the work is the Gene expression datasets which are in its primary forms and they are spliced datasets. As a well known fact these datasets are large in size and mostly they are unstructured in nature. Due to the nature of the datasets, a new classification technique needs to be utilized. Analyzing the possibilities of various classifications and the data preparation technique a new algorithm namely the Enhanced Generalized Sequential pattern algorithm has been devised.

As the data taken for the study is the DNA Splice junctions, necessary preparation and cleaning of data needs to be done. Splice junctions are locations on strings of DNA or RNA where superfluous sections are removed when proteins are created. After the splice, a section, known as the intron, is removed and the remaining sections, known as the exons, are joined together. The dataset consists of sequences of DNA that contain either the part of the DNA retained after splicing, the part that was spliced out or neither. Identifying the sequences of DNA plays a vital role as the earlier machine learning algorithms posed a threat on time of computation. To work with the challenges that have been posed in the earlier methods a new workflow pattern has been devised which not only helps in better cleaning and preparation of data but aids in a better classification method.
The Work Flow

- **Training Data Set**
- **Preprocessing**
- **EGSP**
- **Gene Clustering**
- **Label Assignment**

- **Testing Dataset**
- **Preprocessing**
- **EGSP**

- Modified Naive Bayes Classification

- Classification Result

### A. Candidate Sequence Generation:

The pre-processing is done by the means of Candidate Sequence Generation. The input data is the DNA sequences and the output is the Candidate Set. The procedure for the Candidate Generation is as follows:

The gene sequences are fed into the system. The size of the Gene sequences is taken to be as N. The distance is computed on the gene sequences which extract a Distinct Component from the datasets.

For all the datasets in the study the dataset is broken into individual Gene components. If the distance computed does not contain the existing component it is added into the component list. After the distinct components are extracted a new candidate list needs to be generated. This is achieved by computing the support and the confidence measures. A threshold value is set for the candidates. Based on the set threshold value the candidate sequence is generated.

#### Candidate Sequences Generation

**Procedure:**

1. Read the gene sequences into gene set \( G_N \)
2. Let \( N \) be the size of gene sequences
3. Let \( \text{Comp}_{\text{dist}} \)
   - \( // \) Distinct Component Extraction
   - For \( j=1 \) to \( M \) \( // \) where \( M \) is the number of users
     - For \( i=1 \) to \( N \)
       - Break \( G_{ij} \) into individual gene components
       - Find the distinct components \( C_p \) from \( G_{ij} \)
       - If \( \text{Comp}_{\text{dist}} \) does not contain \( C_p \)
         - Add \( \text{Comp}_{\text{dist}} \leftarrow C_p \)
   - End if
   - End for \( N \)
   - End for \( M \)
4. Overall set of distinct components are extracted.
   - \( // \) Candidate Sequence generation
   - Now generate a list candidate \( (\text{Cd}_{\text{list}}) \) sequences to compute the support and confidence
   - Set \( \text{Cd}_{\text{thres}} \) as the candidate threshold
   - For \( x=1 \) to \( \text{Comp}_{\text{dist}} \)
     - For \( y=1 \) to \( \text{Cd}_{\text{thres}} \)
       - Generate candidate sequence \( (\text{Cn}_{\text{seq}}) \)
       - Updated \( \text{Cd}_{\text{list}} \leftarrow \text{Cn}_{\text{seq}} \)
   - End for \( y \)
   - End for \( x \)

The given dataset is broken into distinct gene components so as to aid in the process of identifying if any characters other than the required EI, IE stops are present. The breaking up of these sequences aids in identifying distinct components, and if the sequence does not contain the required components it is added to compute distance set. This gives a fully extracted distinct component set. This proceeds by the way of calculating the support and confidence measures for the candidate list. In order to attain accuracy over the support and confidence a threshold value is set. Based on the threshold value the candidate sequence is generated. The obtained list is thus updated.

#### B. Enhanced Generalized Sequential Pattern (EGSP)

**Input:** Gene sequences, candidate Set

**Output:** Sequential pattern

**Procedure:**

For \( j=1 \) to \( M \)
For \( i=1 \) to \( N \)
Let \( \text{Seq} \) be the Gene Sequence of the user \( (U_j) \)

Break Seq by \( \text{Cd}_{\text{thres}} \)

\[
\text{Ct} = \begin{cases} \text{Increment Ct} & \text{if Seq} \text{ found in } \text{Cd}_{\text{list}} \\ \text{Ct} & \text{Otherwise} \end{cases}
\]

\[
W_i = \frac{\text{Ct}}{\text{Size} (\text{Cd}_{\text{list}})}
\]

End for \( N \)
End for \( M \)
End for \( f \)
For \( n =1 \) to Number of \( \text{X} \)
For \( M=1 \) to Number of \( \text{X} \)
Let \( \text{Sp}_x \) be the individual Sequence
Let \( \text{P}_{\text{seq}(f)} \) be the \( \text{Cand}_{\text{seq}} \) in \( \text{Seq}_f \)
Calculate occurrence \( \text{G}_{(P)} \) in \( \text{Cd}_{\text{list}} \)
Support \( (\text{Sp}_f) = \text{O}_{(P)} + W_i + W_{ct} \)
End for \( f \)
End for \( n \)
Set threshold \( (\text{Th}_x) \) for Pruning the sequences

Compute min confidence and max confidence of the sequence
\[
\text{Mi}_{\text{cf}} \text{And} \text{Max}_{\text{cf}}
\]
\[
\text{Th}_x = (\text{Th}_x \% \ast \text{Mi}_{\text{cf}}) + \text{Mi}_{\text{cf}}
\]
If \( \text{Conf}_{x} < \text{Th}_x \)

\[
\text{Pr}_x \leftarrow \text{Addseq}_{k} \text{ to Pruned list}
\]
End if

The EGSP algorithm plays an important role which paves the accurate path for classification process. The algorithm takes the input as Gene Sequences and candidate set and the expected output is the sequential pattern. This algorithm works on both the test and the trained data. The first step is the weight estimation where the weight is estimated for all the candidates and all the records. The sequence is broken by candidate threshold. This is computed by the equation
The obtained $Ct$ is divided by the size of the candidate list that was generated in the previous phase. As there are three classes namely El, E2 and N the class weights needs to be generated for each sequence which is given by $W_{cl}$. $X$ is taken as the size for each gene sequence. The no. of sequences available needs to be computed which is given by $Y$. The individual sequence is given by $Seq_i$. $P_{seq_i}$ is taken as the candidate sequence in $Seq_i$. The occurrence of the sequence in the candidate list is calculated using the support factor. The support factor is calculated using the formula

$$Sp_{i} = \sigma_{i}(P_{i}) + W_{cl}$$

Where $Wt$ is the weight of the candidates and $Wct$ is the class weight. Based on the support value achieved the next task is to compute the confidence and Lift. The confidence is computed using the equation

$$\text{Conf}_{i} = \frac{\sum_{i=1}^{N}(Sp_{x} + Sp_{y})}{Sp_{x}}$$

Where $Sp_{x}$ is the support count of individual gene sequence and $Spy$ is the support count of the taken gene sequence. To have more accuracy over the extracted data the lift measure is calculated using the equation

$$\text{Lift}_{x} = \frac{\sum_{i=1}^{N}(Sp_{x} + Sp_{y})}{(Sp_{x} + Sp_{y})}$$

After finding the confidence and support values for the sequences the data needs to be pruned as it has some irrelevant data. A threshold value is set for pruning the sequence. Based on the threshold the minimum confidence and maximum confidence of the sequence is computed. The threshold is set using the formula

$$\text{Th}_{3} = (\text{Th}_{2} * \% \text{Mi}_{ct}) + \text{Mi}_{cf}$$

If the confidence value for the given sequence is lesser than the threshold value the sequence is added into the pruned list.

Now that the data is cleaned and prepared the next process is clustering. The choice of choosing the clustering algorithm is to enhance the classification process and the speed of classification.

### Gene Clustering Algorithm

**Input:** Extracted sequential patterns ($Seq_{i}$)  
**Output:** Clustered Sequences

**Procedure:**

Let $C$ be the number of clusters  
Let $Ch_{iH}$ be the initial cluster head which are set to the clusters

$$\text{count}_{i} = \text{size} (Seq_{i})$$

Add $Mx_{ct}$ to $Ch_{iH}$ // cluster head list  
Compute $Av_{ct} = Mx_{ct} - Mx_{ct} * (Av_{th})$  
Add $Av_{ct}$ to $Ch_{iH}$ // cluster head list  
Compute $Mx_{ct} = Mx_{ct} - Mx_{ct} * (Av_{th})$  
Add $Mx_{ct}$ to $Ch_{iH}$

For $i=1$ to $C$ // $i$ is the number of sequences

For $j=1$ to $J$ // $J$ is the number of $Ch_{iH}$ Size

$$D_{ij} \leftarrow \text{Distance} (\text{Conf}_{iH}, \text{Conf}_{j})$$

$1 \rightarrow \text{index} (\text{mini}(D_{ij}))$

Update $Seq_{i}$ to $ch_{iH}$

End for $j$

End for $i$

The input for the Gene Clustering algorithm is the sequential patterns which have been extracted as an output of the EGSP algorithm. The desired output is the Clustered sequence. $C$ is taken to be the number of Clusters and $ChH$ is the initial cluster head which are set to the clusters. The cluster head selection is done by counting the sequences. The maximum count and the minimum count of the sequences are taken as $Mx_{ct}$ and $Mx_{ct}$, respectively. The maximum count $Mx_{ct}$ is added to the cluster head list $ChH$. Based on the maximum count the average count is calculated. In the same manner minimum count is computed with respect to the maximum and the minimum count. This minimum count is added to the cluster head list. The frequent and the infrequent sequences are updated to the cluster head list. For this computation the distance is calculated based on the confidence of $i$ and $j$ values. The minimum index value is updated to the sequence list, which is then updated to the cluster head list for each and every candidate list $j$.

These clustered sequences are then classified using the Modified Naïve Bayes Classification. 

#### Modified Naïve Bayes Classification

**Input:** Clustered Sequences  
**Output:** Classified output

**Procedure:**

$S$ be the Size of the Total sequences to be classified  
Let $F_{s}$ be the feature set extracted  
Let $L_{b}$ be the labels corresponding to the selected features  
Let $N$ be the number of classes to be identified

For $i=1$ to $N$

$$\text{cnt}_{i} = \text{count} (F_{s}) \text{ in belonging to class cl}_{i}$$

End

Compute Total count as $\text{Cnt}_{T} = \sum_{i=1}^{N} \text{cnt}_{i}$

Compute probabilistic Components for each class as

For $i=1$ to $N$

$$P_{\text{comp}} (i) = \text{cnt}_{i} / \text{Cnt}_{T}$$

End

Let $sd$ be the minimal standard deviation rate  
For each test data in testing Set size ($ts_{s}$)

For each class $cl$

For $M=1$ to number of features ($F_{i}$)  
Compute the probabilistic Components of the feature in each class $cl$

$$P_{x} \leftarrow \text{Get number of data in range of}$$

$$F_{i} - sd \geq \text{data and} F_{i} + sd \leq \text{data}$$

$$P_{t} \leftarrow \text{Get total number of features}$$

End for $m$

$$P_{x}(F_{i}) = \left[ \frac{\text{size} (F_{i})}{P_{t}} \right] \text{ for each class in cl}_{i}$$

Now $P_{x}(cl_{i}) = P_{x}(F_{i}) * P_{\text{comp}} (i)$

End for each class

Let $N$ be three classes

If ($P_{x}(cl_{1}) > P_{x}(cl_{2})$ and $P_{x}(cl_{1}) > P_{x}(cl_{3})$)

Set class label to $cl_{1}$

Else if ($P_{x}(cl_{1}) > P_{x}(cl_{2})$ and $P_{x}(cl_{1}) > P_{x}(cl_{3})$)

Set class label to $cl_{2}$

Else if ($P_{x}(cl_{1}) > P_{x}(cl_{2})$ and $P_{x}(cl_{1}) > P_{x}(cl_{3})$)

Set class label to $cl_{3}$

End if
The modified Naïve Bayes classification aims at classifying the instance, where the clustered sequences are given as input. The size of the total No. of sequences is taken as S. The training set and the Test set sizes are initialized which is given by \((t_r)\) and \((t_s)\) respectively. The No. of classes to be identified is given by N. For each of the classes the count of the feature set is updated into cnti, where i specify which class it belongs to. The total count is then computed. The probabilistic component computation is performed for each class. The main feature of MNBC is that it proceeds by the way of finding the minimum standard deviation for each range of instance. Within each data of the testing set and within each class, the algorithm is said to get the set of attributes for each sequence.

The number of data in a given range is assigned to \(P_{xi}\) if the \(F_i - sd \geq data\) and \(F_i + sd \leq data\). The total no. of features obtained is given by \(P_i\).

The probability for each feature extracted within each class is given by

\[
P_x(F_i) = \prod_{i=1}^{size(F_i)} P_{xi}/P_i
\]

The probability that a given sequence falls within a class is given by

\[
P_x(cl_i) = P_x(F_i) * P_{comp}(i)
\]

Let the three classes be given as cl1, cl2 and cl3 respectively.

If the probability of Class 1 is greater than the probability

class 2 and class 3 the class label is set to cl1. Otherwise if

the probability of class2 is greater than the probability of

class1 and class3, then the class label is set to 2, else it is set
to 3.

Even though the consequences are highly correlated the

MNBC offers a greater improvement over the Naïve Bayes

Classification and the Naïve Hubness Classifier methods.

This is in accordance with the results that have been achieved, though the amount of data taken for the analysis is highly scalable.

III. RESULTS AND DISCUSSIONS

The dataset that has been taken for the work is the Gene

sequence dataset for the Species Homo Sapiens. The data in

its raw form cannot be used as it needs to spliced and

the necessary cleaning and preparation of data needs to be

considered for the work. After the preparation we achieve at

a dataset which is spliced. The Gene expression datasets varies from stage to stage in a human body. Splice

junctions are locations on strings of DNA or RNA where

superfluous sections are removed when proteins are created.

After the splice, a section, known as the intron, is removed

and the remaining sections, known as the exons, are joined
together. Analyzing these sections of data is highly time

consuming. There are many Machine learning algorithms

but the methodology proposed in this paper proves to have a

higher efficiency and accuracy over the existing classification algorithms. The dataset considered for this

work consists of sequences of DNA that contain either the

part of the DNA retained after splicing, the part that was

spliced out or neither. The main focus is to distinguish and

classify them to which case it belongs to.

There are three classes of data namely

- EI
- IE
- N

EI is the Exon –Intron Boundary, IE is the Intron – Exon

Boundary and N belongs to neither of the classes. EI are
called as donors and IE are called as acceptors. An EI

boundary is defined by a short sequence around the

boundary and the absence of a “stop” codon on the Exon

side of the boundary. The Introns are un-translated

intervening sequences in mRNA. The biological

confirmation of prediction is almost always necessary. With

so many genomes being sequenced, it is important to be

able to identify genes and the signals within and around
genes computationally.

A Sample of the datasets that was taken for the study is

given below. The Analysis has been made for three
different sample sizes and the study was carried out for

various measures like F1- Score for Donor, Prediction

Accuracy, and Accuracy for hc19, Accuracy for Hg-38 and

Accuracy factor over different sample sizes.

Once the dataset has been loaded it gets ready for the data

preparation step. The Preparation of the data is based on the

EI , IE stops. The initial dataset consists of both translated

and un-translated datasets, applying proper methods they

are translated. But still the dataset seems to contain some

records which are neither of the EI or IE categories. Analyzing

and fixing them into the reliable category is the task.

After the preparation of the data the candidate support and

classification aims at classifying

candidate weight of the input data. To further enhance and

improve the process the confidence and lift are measured

for each sequence with its support value. Once the

confidence and lift measures have been computed, the

sequences needs to be pruned as it may contain IE stops.

We are concerned in only the EI codons and not IE codons.
The output gives us the Extracted Sequential pattern from

the frequent sequences.

The extracted patterns then need to be clustered. The

average, minimum and maximum support count is

calculated for each clusters and added to the cluster head.

To find out whether the sequences fall into the specified

cluster head, the distance is calculated based on the

confidence measures which were arrived earlier. The

minimum distance tells us that the given sequence falls into

one of the cluster heads. These clustered sequences needs to

be classified using the Modified Naïve Bayesian

classification algorithm. The inputs taken are the clustered

sequences, testing data and the training data. The features

need to be extracted from these sequences and a list of

feature set is created. The data set consist of three classes ,

where some of the records are unlabeled , these unlabelled

classes are categorized as N as discussed earlier and these

datasets needs to fixed either into the EI or the IE classes.

The probabilistic components for each of these classes are

computed. To restrict the error values and identify them

accurately a minimal standard deviation is computed for
each class. The probability component for each feature is

computed using the formula
\[ \text{Px}(F_i) = \prod_{i=1}^{\text{size}} \left( \frac{F_i}{P_i} \right) \]  

The normal Bayesian classification follows by the way of summing up the probabilistic values, but in this proposed work a product of these probabilities is computed as the data obtained from the previous algorithm was of the clustered format. Based on the computed probability the components need to be fixed into the respective classes. The comparison of values is done with all the three classes and the new labels are set. Thus the classification has been arrived efficiently using the EGSP and MNBC algorithms. The efficiency and the accuracy measures are supported by the way of performing various statistical tests on the existing methods and the proposed methods. The F1 score is used to give a weighted average on the precision and recall considering the false positives and the false negatives. Though an uneven class distribution was observed the F1 score has given a higher performance over the existing methods. The table below shows the comparison of F1 score on various existing methods and the proposed methods.

Table 1
CLASSIFICATION PERFORMANCE BASED ON F1 SCORE

| Algorithms     | F1-Score |
|----------------|----------|
| DBN            | 0.846    |
| SVM(RBF)       | 0.774    |
| SVM(Sigma)     | 0.714    |
| Gene splicer   | 0.747    |
| Splice Machine | 0.775    |
| Proposed       | 0.981    |

The classification performances have been compared using various classifiers and is compared with the proposed methods. As can be seen clearly from Table 1 of values and the graph it is observed that the proposed methodology has gained in its F1 score value well above the existing methods like DBN, SVM(RBF), SVM(Sigma), Gene Splicer, Splice Machine. This gives us the confidentiality measure on the proposed method and sure is a cliché to the classification method.

Applying this method further enhances that it is certain to touch the mark of 1 on the F1 score values. Taking into consideration of how DBN works with, the datasets could be trained greedily one layer at a time. Though it proves to give a better classification it has been observed that it works well with unsupervised data. But the dataset that has been taken for the study comprises of supervised, semi-supervised and un-supervised. Comparatively from the Table 1 we could observe that the value obtained from DBN is 0.846 and that from the MNBC is 0.981. All the other classifiers like the SVM (RBF), SVM (sigma), Gene Splicer and Splice Machine it has been observed from the graph that surely there is a dip in the performance comparing to the proposed classifier. This is achieved by computing the minimum standard deviation for the each cluster head and this distance measure helps in computing the probabilistic components for each class. This probabilistic component is computed by using the formula

\[ \text{Px}(F_i) = \prod_{i=1}^{\text{size}} \left( \frac{F_i}{P_i} \right) \]

The output achieved gives us a view over the better classifier. This shows a clear demarcation that the proposed classifier stands out from all the other classifiers.

Table 2
Prediction accuracy of classifiers

| Dataset     | Accuracy Prediction |
|-------------|---------------------|
| Splice      | Min      | Max      | Average  |
| Accuracy in 12 | 95.87    | 96.73    | 96.18    |
| Accuracy in 2  | 96.65    | 96.93    | 96.81    |
| Proposed     | 95.68    | 98.88    | 97.08    |

Figure 1
Classification Performances using F1 score

The classification performances have been compared using various classifiers and is compared with the proposed methods. As can be seen clearly from Table 1 of values and the graph it is observed that the proposed methodology has gained in its F1 score value well above the existing methods like DBN, SVM(RBF), SVM(Sigma), Gene Splicer, Splice Machine. This gives us the confidentiality measure on the proposed method and sure is a cliché to the classification method.

Applying this method further enhances that it is certain to touch the mark of 1 on the F1 score values. Taking into consideration of how DBN works with, the datasets could be trained greedily one layer at a time. Though it proves to give a better classification it has been observed that it works well with unsupervised data. But the dataset that has been taken for the study comprises of supervised, semi-supervised and un-supervised. Comparatively from the Table 1 we could observe that the value obtained from DBN is 0.846 and that from the MNBC is 0.981. All the other classifiers like the SVM (RBF), SVM (sigma), Gene Splicer and Splice Machine it has been observed from the graph that surely there is a dip in the performance comparing to the proposed classifier. This is achieved by computing the minimum standard deviation for the each cluster head and this distance measure helps in computing the probabilistic components for each class. This probabilistic component is computed by using the formula

\[ \text{Px}(F_i) = \prod_{i=1}^{\text{size}} \left( \frac{F_i}{P_i} \right) \]

The output achieved gives us a view over the better classifier. This shows a clear demarcation that the proposed classifier stands out from all the other classifiers.

Figure 2 Accuracy in Prediction

Prediction accuracy plays a vital role in the determination of the data sets belonging to a specific category. The accuracy of the spliced data has been analyzed on three scales namely the Minimum, Maximum and the average. As the prepared data is spliced in nature, three comparisons were done based on Accuracy in 12, Accuracy in 2 and the proposed method. The proposed method shows a clear gain over the two existing methods. On all the three measures the proposed method has achieved a good minimum, a good maximum and a good average score.

Though it has been a proven fact that in the field of study of pulmonary diseases 12 sigma played a significant role it could be observed form Table 2 that using minimum factor for prediction 12 sigma has a
An Enhanced Generalized Sequential Pattern Classification For Sequence Datasets

The accuracy has been proved for Classification and prediction. As there has been challenges in the efficiency factor over scalability of the dataset a comparison has been made by gradually increasing the no. of samples from 12000 to 60,000. It is very obvious from the chart that the proposed method has a clear leverage over the existing sampling methods.

IV CONCLUSION

The EGSP and the MNBC algorithms were proposed to have an edge over the existing classification algorithms, where the existing algorithms proved to work well for smaller samples. But when it comes to the point of scalability they have however a shown a clear descent in the values achieved. The threshold values that have been set for the candidate generation offered a way to attain accuracy for the further processing. The pruning of the sequences have contributed to the fact for decidability factor by means of computing the minimum and the maximum confidence measures. The clustering algorithm provided a means of speeding the process of classification by calculating the distance between the confidence measures of the sequence to the cluster head. These clusters have apparently contributed to the faster and accurate classification. The Naive Bayesian classification has been modified in the way of working by calculating the standard deviation measures and with the standard deviation achieved the probability of the data belonging to a particular class has been achieved. The statistical measures have undoubtedly proved that the proposed algorithms stand a way ahead of all the existing methodologies.

REFERENCES

1. Devi Arockia Vanitha ,Devaraj D, Venkatesulu, “Gene Expression Data classification using support Vector Machine and Mutual Information–based Gene selection”, Procedia Computer science 47(2015)13-21.
2. Heba Abusamra, “A comparative study of feature selection and classification methods for gene expression data of glioma”, Procedia Science Direct, Elsevier Issue:10.1016/j.procs.2013.10.003. For Conference.

By calculating the standard deviation measures the clustering algorithm provided a means of computing the minimum and the maximum confidence measures. The clustering algorithm provided a means of speeding the process of classification by calculating the distance between the confidence measures of the sequence to the cluster head. These clusters have apparently contributed to the faster and accurate classification. The Naive Bayesian classification has been modified in the way of working by calculating the standard deviation measures and with the standard deviation achieved the probability of the data belonging to a particular class has been achieved. The statistical measures have undoubtedly proved that the proposed algorithms stand a way ahead of all the existing methodologies.

REFERENCES

1. Devi Arockia Vanitha ,Devaraj D, Venkatesulu, “Gene Expression Data classification using support Vector Machine and Mutual Information–based Gene selection”, Procedia Computer science 47(2015)13-21.
2. Heba Abusamra, “A comparative study of feature selection and classification methods for gene expression data of glioma”, Procedia Science Direct, Elsevier Issue:10.1016/j.procs.2013.10.003. For Conference.

The accuracy has been proved for Classification and prediction. As there has been challenges in the efficiency factor over scalability of the dataset a comparison has been made by gradually increasing the no. of samples from 12000 to 60,000. It is very obvious from the chart that the proposed method has a clear leverage over the existing sampling methods.

IV CONCLUSION

The EGSP and the MNBC algorithms were proposed to have an edge over the existing classification algorithms, where the existing algorithms proved to work well for smaller samples. But when it comes to the point of scalability they have however a shown a clear descent in the values achieved. The threshold values that have been set for the candidate generation offered a way to attain accuracy for the further processing. The pruning of the sequences have contributed to the fact for decidability factor by means of computing the minimum and the maximum confidence measures. The clustering algorithm provided a means of speeding the process of classification by calculating the distance between the confidence measures of the sequence to the cluster head. These clusters have apparently contributed to the faster and accurate classification. The Naive Bayesian classification has been modified in the way of working by calculating the standard deviation measures and with the standard deviation achieved the probability of the data belonging to a particular class has been achieved. The statistical measures have undoubtedly proved that the proposed algorithms stand a way ahead of all the existing methodologies.

REFERENCES

1. Devi Arockia Vanitha ,Devaraj D, Venkatesulu, “Gene Expression Data classification using support Vector Machine and Mutual Information–based Gene selection”, Procedia Computer science 47(2015)13-21.
2. Heba Abusamra, “A comparative study of feature selection and classification methods for gene expression data of glioma”, Procedia Science Direct, Elsevier Issue:10.1016/j.procs.2013.10.003. For Conference.

Table 3 Accuracy Vs Size of the samples

| Method     | No of Samples |
|------------|---------------|
|            | 12000 | 25000 | 60000 |
| CD         | 94.49 | 94.36 | 98.17 |
| Persistent CD | 45.58 | 46.46 | 98.68 |
| Parallel Tempering | 95.84 | 95.74 | 98.52 |
| Proposed   | 95.68 | 96.7  | 98.88 |

| Accuracy vs Size |
|------------------|
| No.of.Samples    |
|                  |
| 12000            |
| 25000            |
| 60000            |

3. Hung-Yi Lin, “Gene Discretization based on EM clustering and adaptive sequential forward gene selection for molecular classification”, Elsevier, Applied Soft Computing 48(2016) 683-690.
4. Jesus Maillo,Sergio Ramirez,Isaac Triguero,Francisco Herrera, “KNNSIS: An interactive Spark-based design of the K-nearest Neighbors classifiers for big data” Knowledge based Systems, Elsevier 000(2016)1-13.
5. Jia Lv,Qinke Peng,Xiao Chen, Zhi Sun , “A multi-objective heuristic algorithm for gene expression microarray data classification”, Elsevier, Expert Systems with Applications 59(2016)13-19.
6. Konstantina Kourou, Costas Papaloukas,Dimitrios I Fotiadis, “Integration of pathway Knowledge and Dynamic Bayesian Networks for the prediction of Oral Cancer Recurrence”, IEEE 2016.
7. Krissitian Buza, “Classification of Gene Expression data: A Hambness-aware semi-supervised approach”, Elsevier, Computer Methods and Programs in Biomedicine 127(2016) 105-113.
8. Pradeep K Sharma, Vaibhav Sharma, Jagrati Nagdiya, “A Proposed Method for Mining High Quality itemset with Transactional weighted utility using Genetic algorithm Technique”, IJRCSE, Vol -5, Issue 1, pp 31-35, 2017.
9. Sara Tarek, Reda Abd Elwahab, Mahmoud Shoman, “Gene Expression based cancer classification”, Egyptian Informatics Journal 2016.
10. T. Semhimbelvi, R.Parimala, “Improving Clustering Accuracy using Feature Extraction Method "”, IJRCSE, Vol-6, Issue-2, pp 15-19, 2108.
11. Thanh Nguyen, Saeid Nahavandi, “Modified AHP for Gene Selection and Cancer Classification using Type-2 Fuzzy Logic”, IEEE Transactions on Fuzzy Systems, Vol 24 No.2 April 2016.

AUTHORS PROFILE

Immaculate Mercy A is pursuing her Ph.D in Computer Science from Bharathidasan University, Tiruchirapalli. She holds a Masters Degree in Computer Applications, Masters in Computer Science Engineering and MPhil in Computer Science. She has over 16 years of experience in teaching at the Under Graduate and the Post Graduate level and 8 years of industry experience, which involved in software development and e-content writing. She also has a proven experience in Corporate Training. Her research work focuses on Data Science, Prediction Analysis, Data Mining, Big Data Analytics and Cloud Computing.

Retrieval Number: A4895119119/2019@BEIESP
DOI: 10.35940/ijitee.A4895.119119

Published By:
Blue Eyes Intelligence Engineering & Sciences Publication

2528
Chidamabaram M pursued his Masters in Computer Science and M.Phil from Bharathidasan University, M.B.A from Periyar University and Ph.D from Vinayaka Mission’s. in Computer Science. He has over 21 years of experience in teaching at the undergraduate and post Graduate level. He has guided over 25 scholars towards M.Phil degree and 8 scholars towards Ph.D degree. He has published more than 40 research papers in various National and International journals, 8 conference papers. He is working as an Assistant Professor in Computer science in RSGC, Thanjavur. His areas of Research are Cloud, Grid Computing, Data Mining and Big Data Analytics.