Efficient Scheme for Classifying Grass Genomes

S.S.Patil, B.V.Dhandra , U.B.Angadi

Abstract- **Classifying genome sequence data for predicting homology and properties of genome sequences is a challenging task due to its large size. Clustering genome data sequences reduces data size and simplifies the classification process by reducing training time. Classification of grass genome sequences, the largest genomes dataset available in the public domain, is best suited for developing strategy for innovative prediction of structure and function of DNA sequences. Clustering techniques are used to generate an efficient set of prototypes to decrease the classification time without compromising on Classification Accuracy (CA). This study describes the use of data mining algorithms with multiple distance measures – (1) Leader algorithm using Euclidean distance to generate non-overlap motifs with local alignment as the features of the genome sequences, (2) Hamming Distance Classifier (HDC) and Nearest Neighbor Classifier (NNC) algorithm to generate a set of prototypes to increase the CA. We have observed that the Nearest Neighbor Classifier modified with hamming distance increases the CA and reduces the classification time. The clusters generated using the HDC with the Minimum Distance Classifier (MDC) was found to optimize the requirements of computation time and space. Experimental data was further compared using Leader Based Nearest Neighbor Classifier (LBNNC) algorithm with global alignment technique. These results suggest that the Nearest Neighbor Classifier algorithm provides optimal CA in both training and test phases but takes longer duration. The best time complexity was achieved in MDC and LBNNC. Using HDC we achieved 83% in actual classification of the grass genome data.**

Keywords: Classification Accuracy (CA), Hamming Distance Classifier (HDC), Nearest Neighbor Classifier (NNC), Minimum Distance Classifier (MDC), Leader Based Nearest Neighbor Classifier (LBNNC)

I. INTRODUCTION

Clustering large datasets is a challenging task in data-mining. Sequence clustering is very important to study genome data [Vijaya et.al. 2006, Luo et.al.2002], which consists of a long stretch of nucleotide sequences with three dimensional folding. We propose schemes using motif based clustering to analyze grass genome sequences data [Alison 1999, Gaut 2002]. Clusters of genome sequences are made through sequence alignment and pattern recognition. Classification is performed using similarity or distance measures [Needleman and Wunsch.1970]. The evaluation of classifiers is assessed on CA: Based on training and test data

$$CA = \frac{\text{Number of test data cluster classified correctly}}{\text{Total no of test data}} \times 100$$

new sequences can be classified using sequence similarity to the known class/family of the sequences [Fischer and Paterson 1974]. This helps in predicting the structure and function of unknown sequences to save the expenses on the biological experiments. Classification of multiple sequences is important to determine the molecular evolution of the gene family and to understand their current status of evolution in biological systems. Comparing the whole length of sequences with each other using distance measure is very difficult. A sub string of sequence motifs can be used to generate genome profiles. The grass family (grasses cover >20% of the earth's land surface, often dominate temperate and tropical habitats) includes more than 10,000 species of plants. Although, it is a relatively small family compared with flowering plants, it surpasses others ecologically and contributes to economic growth (maize, wheat and rice are the staple food crops) the world over [Gaut 2002]. Evolution wise, the grass family is fascinating due to the degree of variation found in genome size; ploidy level and chromosome number [Wang 2005, Yu, et al. 2005]. Most of the comparative genomic studies have been initiated in barley, wheat, maize, rice, and sorghum to understand the diversity and structure function relationship of the genomes [Wei, et al.2007, Wang 2005]. Our paper reports on the classifier algorithm based on clustering in Pattern Recognition system that uses memory and space efficiently.

II. RELATED WORK

Global Alignment [Dayhoff and Schwartz 1979, Needleman and Wunsch.1970] is used to align the genome sequences. An important component of our work is methodology/algorithm for alignment based on similarity. The general practice is to carry out the alignment of sequences in a cluster by inserting "gap" ("gap" helps to align two sequences by inserting some gap at different locations. Whatever is the unknown is the score measurement). Sequence homology is determined by match award, mismatch and gap penalty. Local alignment system [Smith and Waterman 1981] aligns the motifs at multiple locations in sequences instead of aligning entire sequences [Felsenstein, et al. 1983]. One of the algorithms for local alignment is given in [Smith and Waterman 1981]. Time complexity of the algorithm is O(n^2l^2), where "l" is dimension of motif, "n" is number

S.S.Patil is with Dept. Computer Science, University of Agricultural Sciences, Bangalore, India. spatils@yahoo.co.uk ,

B.V.Dhandra is with Dept. Computer Science, Gulbarga University , Gulbarga, India. bv_dhandra@yahoomail.com

U.B.Angadi ARIS is with National Institute of Animal Nutrition and Physiology, Bangalore, India ubangadi@gmail.com

of sequences. Dynamic algorithm is widely used in aligning two sequences where time and space complexities are O (mn), where "m" is length and "n" is a pair of sequences. When we need only similarity, space complexity reduces to O (n). In this paper, we use a modified dynamic programming technique which is used for computing only similarity measures with local alignment [Smith and Waterman 1981].

Terminology

Given a set Q of genome sequences, S1, S2, S3,…, S_n of varying length L1, L2, L3,... ,Ln, each of these sequences can be presented as Si=ai1, ai2, …..,aiL Є α for i =1…n, where α is a set of DNA sequences. The common motifs from the set of sequences with a similar measure can be generated based on threshold using Smith and Waterman's local alignment technique. It provides the frequency table of size, "n x m" based on the number of occurrence of segments, where 'n' is the number of sequences and 'm' is the number of motifs. The calculated distance between motifs (using frequency table) is used in clustering the sequences based on user specified threshold.

III. THE CLASSIFICATION ALGORITHMS

In this study, four classification algorithms are used.

A. NNC:

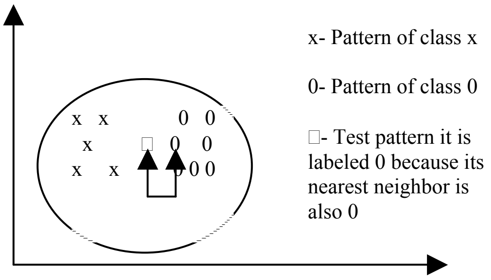
x- Pattern of class x
0- Pattern of class 0
□- Test pattern it is labeled 0 because its nearest neighbor is also 0
x x 0 0
x □ 0 0
x x 0 0

Fig-1: Nearest Neighbor Classifier (NNC)

NNC [Cover and Hart, 1967] is used in [Yi and Eric, 1993; Salzberg and Cost, 1992] genome secondary structure predictions. NNC is computationally very expensive for large data sets [Needleman and Wunsch, 1970] which classify the new sequence by using blocks of genome sequence segments [Mohseni, et. al., 2004]. The test sequence is compared with these blocks and the local or global similarity score is determined. This system does not consider the gaps that are conserved in a set of multiple aligned sequences, and group as the set of genome sequences exhibit sequence similarity [Dayhoff and Schwartz, 1979]. Both the test and training sequence is classified to the group for which this score value is maximized in terms of CA. They have tested their method on the PROSITE database consisting of known genome sequences. Time and space complexity of pairwise local alignment algorithm is O(*uv*), therefore the time complexity to classify a new sequence using NNC is O((uv)n), where u and v are length of the two sequences, where n is the total number of sequences in training set.

B. MDC:

Minimum (mean) Distance Classifier (MDC) is applied [Duda et.al. 2000] for numerical data sets wherein centroid would be the class representative. We used the hamming distance results for the given weightage and the sum of similarity scores suitable for genome sequences in our analysis, since centroid cannot be defined for a group of sequences.

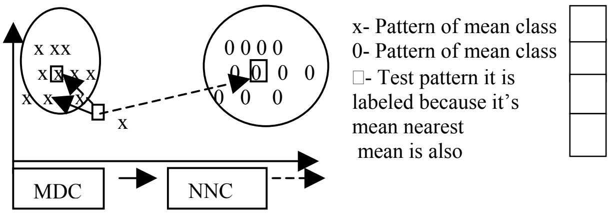

Fig-2: Minimum (Mean) Distance Classifier (MDC)

```
1. Initialize motifs say first segment, as the representative of the first motif, add it to motifs list, and set segmenting counter.
2 If (distance with motif1, substring2 less than or equal to threshold) then
        motif1=substring2
        else
        new motif2=substring2
          Assign it to Class-i; flag = S1
    Class-1 = S1
    Call (hamming distance)
    Hamming distance= Sum of difference
     for all patterns, i=1 to n
If (hamming distance with S1 and S2 ≤ threshold) then
                    Class-1= S2
         else
                     new Class2  = S2
         // end if
    end of loop
         k=1; (N= No. of Sequences, M=No. of motifs)
    for i=1 to N
    for j=1 to M
         md(i, k) = (motif(i,j)+ motif (i, j+1) + motif (i, j+2))/k+1;
         Q=K ,   Q is Total number of Minimum distance classifiers
    end of loop.
```

Fig-3: MDC Algorithm

C. LBNNC:

Leader is an incremental clustering algorithm, in which each cluster is represented by a leader i, for i=1...k, k is the number of clusters generated using a suitable threshold. In this, first sequence is selected as leader of the first cluster and categorization of the remaining sequences will be made either in existing clusters or choosing them as leader of a new cluster based on similarity and threshold.

D. HDC

```
Input: Si – A set of N sequences for i 1 to N, t-Threshold
Output: Cj: A set n classes for j 1 to n,    Lj: A set n leaders for j 1 to n
Initialize:
j=1; Cj = Si ;          (S1 is member in C1)
Lj = Sj           (L1 is leader of C1)
j++
for i=2 to N
     finding nearest cluster for Si using
     threshold t call( hamming distance)
 if nearest  exists
   Si is member of Cj
 else
  create new class Cj+1
  Si  is  member of Cj+1
   Lj+1=Ci
 end if
end for
```

Fig 4: HDC Algorithm

To calculate the distance between the phylogenetic profiles of two genome sequences, we used Hamming distance. To determine the Hamming distance between two genome sequences, we sum up the number of times the genome motif is found, $D_H = \sum_{i=1}^{n} d_i$, where *n* is the number of genome motifs, $d_i = 0$ if the orthologs of genome sequences are either present or both absent in genome motif i, and $d_i = 1$ otherwise. The above function will compute the Hamming distance of two integers (considered as binary values, that is, as sequences of bits). The running time of this procedure is proportional to the Hamming distance rather than the number of bits in the inputs [Henikoff and Henikoff 1992, Wang, et.al. 2005]. We have used HDC to classify the genome sequences in this study.

IV. EXPERIMENTAL RESULTS

Among the four classification schemes used in this work, LBNNC requires cross validation as it is order dependent.

A. Cross validation of LBNNC:

Cross validation is an established technique for estimating accuracy of a classifier and is normally performed either using a number of random test/training partitions of the data or using m-cross fold-validation. We present a technique for calculating the complete cross validation for LBNNC: i.e., averaging overall desired test/train partitions of grass genome data (Table-1)

Table.1: *Cross validation for GGSDs*

GGSD1	Thres hold	Classes	Training time (sec)	Test time (sec)	CA
LBNNC	0.7	4	12	8	99.95
Cross validation for LBNNC	0.7	4	12.2	1.4	100

To evaluate the performance of the algorithms, we consider the Grass Genome data in the FASTA format [Mount 2002, Peter and Rolf 2000]. Grass Genome data set consists of different grass families like Switch grass, Bermuda grass, and Sorghum. Combination of grass genome data sets having 2.64 lakh genome sequences was collected and used in this study. We carried out classification using four schemes: NNC, MDC, LBNNC and HDC for genome sequences classification. The data set makes the divisions of approximately 5000 genome sequences in each set. Out of 5000, we chose 3000 as training data and the remaining2000 sequences as the test data for analysis *(Table-2)*

Table.2: *Time and CA for GGSDs*

Algorithm	Data Set	Training data	Test data	Threshold	Classes	Training time (sec)	Test time (sec)	CA
NNC	1	3000	2000	0.7	23	20	218	99.63
	2	3000	2000	0.7	15	21	247	99.65
	3	3000	2000	0.7	15	22	241	99.89
	4	3000	2000	0.7	10	21	230	99.89
	5	3000	2000	0.7	8	22	252	99.84
MDC	1	3000	2000	0.7	3	15	10	99.8
	2	3000	2000	0.7	5	17	11	99.75
	3	3000	2000	0.7	8	10	8	99.35
	4	3000	2000	0.7	3	6	4	99.65
	5	3000	2000	0.7	3	15	10	99.3
LBNNC	1	3000	2000	0.7	4	12	8	99.95
	2	3000	2000	0.7	3	11	8	99.95
	3	3000	2000	0.7	5	11	8	99.95
	4	3000	2000	0.7	5	20	8	99.25
	5	3000	2000	0.7	4	11	8	99.85
HDC	1	3000	2000	0.7	20	13	25	99.4
	2	3000	2000	0.7	27	14	23	99.35
	3	3000	2000	0.7	22	18	18	99.35
	4	3000	2000	0.7	12	20	15	99.6
	5	3000	2000	0.7	11	16	30	99.65

Table-3*: Comparison of CA with times of training time and test time*

Comparis on with	CA	Test-time	Training time	Fold times with NNC	
				Train time	Test-time
NNC	99.63	218	23	1	1
MDC	99.65	4	6	54.5	3.8
LBNNC	99.95	8	11	27.3	2.1
HDC	99.60	15	20	14.5	1.2

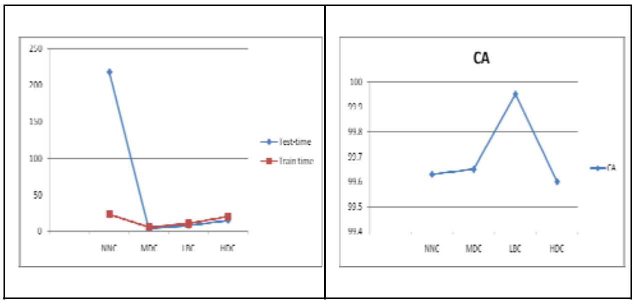

Fig.5. Training Time, Testing Time and CA of schemes

The experimental results are shown as time complexity- in MDC, the best among all in training time (sec) and test time (sec). LBNNC, took second place but stood first in CA as shown in, Table-3

V. DISCUSSION AND CONCLUSION

Clustering by pattern similarity is an interesting and challenging problem. The computational time and space complexity can be high to cluster the genome sequences. We have presented the pattern-based similarity clustering for genome sequences for mining several types of frequent pattern classes in large datasets. Our scheme of classifier performance study shows that the algorithm derived from the pattern-based motifs is good. For our analysis the following algorithms were used to perform clustering and classification of large genome dataset. 1) NNC, 2) MDC, 3) LBNNC, 4) HDC. It is apparent that the CA is almost the same using leader as compared to global alignment and leader with hamming distance. The Generalized Hamming Distance problem has been described and the possible applications are mentioned. The approach to obtain similarities as the inner product of the vectors representing the characters enables us to use the techniques to reduce the cost of computation of similarities, and at the same time, keep the error as low as possible. By taking into account the frequency of characters in the pattern, the errors can be further reduced. The time complexity achieved was the best in MDC and HDC and CA achieved was the best in LBNNC.

REFERENCES

[1] Alison Abbot "Bioinformatics institute plans public database for gene expression data," *Nature,* vol. 398, 1999, pp.638-646.

[2] Cover and Hart, "Show that the error for NN classifier is bounded by twice," *Pattern Recognition Letters*, 1967.

[3] G.C. Aggarwal, C Procopiuc,J. Wolf P.S.Yu and J.S.Park. "Fast algorithm for projected clustering," SIGMOD, 1999.

[4] G.C. Aggarwal, P.S.Yu. "Finding generalized projected cluster in high dimensional spaces," SIGMOD, 2000, pp. 70-81.

[5] Gaut, B.S. "Evolutionary dynamics of grass genomes. New Phytol," vol. 154, 2002, pp. 15–28.

[6] H B. Shen, and K.C. Chou, "Using ensemble classifier to identify membrane protein types," *Amino Acids*, 2007

[7] J. Felsenstein, S. Sawyer, and R. Kochin. "An efficient method for matching nucleic acid sequences: inference and reliability," *Annu. Rev. Ecol. Syst.*, Vol.14, 1983, pp.313-333.

[8] K. Bayer, J Goldstein, R. Ramakrishnan and U Shaft. "When is nearest neighbors meaning full," Proc. of the Int.Conf. Database Theories, 1999, pp. 217-235.

[9] M. Dayhoff and R. Schwartz. "Matrices for detecting distant relationship," *Atlas of Protein Sequences*, 1979, pp. 353-358.

[10] M. Fischer and M. Paterson.," String-matching and other products". Proc. SIAM-AMS Complexity of Computation, 1974, pp. 113-125.

[11] P.A. Vijaya, M.N. Murty and D.K. Subramanian. "An efficient incremental Clustering Algorithms for large data set," Proc. ICAAI, Kolhapur, India, 2005, pp79-85.

[12] Phuongan Dam, Victor Olman, Kyle Harris, Zhengchang Su and Ying Xu, ",Operon prediction using both genome-specific and general genomic information," *Nucleic Acids Research,* Vol. 35, 2007,pp. 288-298.

[13] R Agrawal, J.Gehrke, D.Gunoplus and P Raghavan, "Automatic Space clustering of high dimensional data for data mining applications," SIGMOD, 1998, pp.94-105.

[14] R Luo, Z Feng, J Liu , "Prediction of protein structural class by amino acid and polypeptide composition ," *European Journal of Biochemistry*, 2002

[15] S, Mohseni –Zadeh, P. Brezellec, and J.L.Risel, "Cluster-c an algorithms for the large scale clustering of protein sequences based on the extraction of maximal clique," Computational Biology and chemistry., 2004

[16] S. Henikoff and J. D. Henikoff. "Amino acid substitution matrices from protein blocks". Proce. National Academy of Science USA, Vol. 89(22) 1992, pp.10915-10919.

[17] S.B. Needleman and C.D. Wunsch. "A general method applicable the search for similarities in the amino acid sequences of the proteins", J.*Mol. Biol.* Vol.48, 1970, pp.443-453.

[18] S.S.Patil, M.N.Murty and S.S.Benchalli, "Clustering algorithms of Plant protein sequences," Proc, Intl. Conf. BAAMH, UAS, Bangalore, 2006, pp.168-176.

[19] S.S.Patil, M.N.Murty and S.S.Benchalli, "Clustering of Plant protein sequences," Proc.Natl. Conf. RAFIT, PU Patiala, 2005,pp. 261-264.

[20] Sadeh, I., "Universal data compression algorithm based on approximate string matching, Probability," *Engineering and Informational Sciences*, Vol. 101996, pp.465-486.

[21] Salse, J., Bolot, S., Throude, M., Jouffe, V.,Piegu, B., Quraishi, U.M., Calcagno, T., Cooke, R., Delseny, M., and Feuillet, C. (2008). "Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution." *Plant Cell,* vol. 20pp. 11–24.

[22] T.F. Smith and M.S.Waterman. "Identification of common molecular subsequences," J.*Mol. Biol*., vol.147, 1981, pp.195-197.

[23] U.B.Angadi, M.Venkteshalu, S.S.Patil and P.Abraham Prabhu. "Segment based technique for classification of large set of DNA/Protein sequence data," Nat. Conf. AAT, Kailaslingam Univerisy 2008.

[24] Wang, X., Shi, X., Hao, B., Ge, S., and Luo, J. "Duplication and DNA segmental loss in the rice genome: implications for diploidization," *New Phytol.* Vol.165, 2005, pp. 937–946.

[25] Wei, F., et al. "Physical and genetic structure of the maize genome reflects its complex evolutionary history," *PLoS Genet.* Vol.3, 2007, p.123.

[26] Yu, J., et al. "The genomes of Oryza sativa: A history of duplications," *PLoS Biol*, ,vol. 3 2005,p. 266

[27] Z Wang, "Spectral Analysis of Protein Sequences," 2005 lib.ncsu.edu