Clustering Algorithms for ITS Sequence Data with Alignment Metrics

Andrei Kelarev[1], Byeong Kang[1], and Dorothy Steane[2]

[1] School of Computing, University of Tasmania
Private Bag 100, Hobart, Tasmania 7001, Australia
{andrei.kelarev, bhkang}@utas.edu.au
http://www.comp.utas.edu.au/users/kelarev/
http://www.comp.utas.edu.au/users/bhkang/
[2] CRC Forestry and School of Plant Science
Private Bag 55, University of Tasmania, Hobart, Tasmania 7001, Australia
dorothy.steane@utas.edu.au
http://www.crcforestry.com.au
http://fcms.its.utas.edu.au/scieng/plantsci/

Abstract. The article describes two new clustering algorithms for DNA nucleotide sequences, summarizes the results of experimental analysis of performance of these algorithms for an ITS-sequence data set, and compares the results with known biologically significant clusters of this data set. It is shown that both algorithms are efficient and can be used in practice.

1 Introduction

The investigation of DNA data sets has broad applications in medical and health informatics and many branches of biology. Data mining and machine learning have been crucial in the development of these research areas (let us refer, for example, to Gedeon and Fung [3], Kang, Hoffman, Yamaguchi and Yeap [6], Li, Yang and Tan [10], Webb and Yu [13], Zhang, Guesgen and Yeap [17], Zhang and Jarvis [18]).

The present paper describes and investigates two clustering algorithms that can be used to group a data set of DNA sequences into clusters. In order to achieve significant correlation between clusterings produced by these machine learning algorithms and biological classifications, we have relied on measures of strong similarity between sequences. Such measures have not been considered for these algorithms before. Here we present the results of an experimental analysis of the performance of our new algorithms using a data set derived from the internal transcribed spacer (ITS) regions of the nuclear ribosomal DNA in *Eucalyptus*, and compare the results with clusterings published by Steane *et al.* [12].

For preliminaries on DNA molecules we refer to the monographs Baldi and Brunak [1], Durbin, Eddy, Krogh and Mitchison [2], Jones and Pevzner [5] and Mount [11]. Background information on clustering algorithms can be found in Witten and Frank [15].

A. Sattar and B.H. Kang (Eds.): AI 2006, LNAI 4304, pp. 1027 – 1031, 2006.

2 Clusters of the k-Means Algorithm with Alignment Metrics

Both our algorithms use highly biologically significant alignment scores as a metric and obtain significant results. The novel character of our method is in using a sophisticated and highly informative distance metric based on alignment scores well known in bioinformatics. Every alignment produces an alignment score that measures the similarity of the nucleotide or amino acid sequences.

The alignment scores in our algorithms provide an accurate measure of similarity that is more significant biologically. Alignment scores have properties that differ from those of the Euclidean metrics and their simple modifications discussed, for example, by Witten and Frank [15] (Section 6.4). Hence our algorithms have to be designed differently. First, it is impossible to do simple calculations for the alignment score metric involved in various Euclidean clustering algorithms. For example, it is impossible to find a DNA sequence that is the "midpoint" or "mean" of two DNA sequences. Secondly, it is important to minimize the number of distance calculations, because each of them is time consuming.

Our first algorithm is a new version of the k-means clustering algorithm. The traditional *k-means* algorithm implemented in WEKA environment uses standard Euclidean distances (Witten and Frank [15], Section 4.8). Our algorithm uses the metric of alignment scores to establish similarity between sequences. The unusual character of this metric prohibits direct computation of the mean of a set of sequences in a cluster.

In order to make the algorithm faster, it is desirable to minimize the number of times the alignment scores have to be found. This is why the algorithm operates on the set of given sequences only and does not create any new sequences as means of the given ones. Every alignment score between each pair of the given sequences is found once during a pre-processing stage of the algorithms, and then these scores are looked up in a table during the process of looking for clusters.

The initialization stage proceeds as usual: k sequences are randomly chosen as centroids of clusters, and every other sequence is assigned to the cluster of its nearest centroid. Every iteration of the algorithm looks at each current cluster in turn. It then analyses all sequences of the cluster trying to determine which of them would be the best centroid for this particular cluster. Suppose that the algorithm is considering a cluster C. As a new centroid for this cluster our algorithm is going to choose the sequence s in C with the property that the sum of all distances from s to all other sequences in C is minimal. This is in fact precisely the property of the standard Euclidean means that is essential for the operation of the traditional k-means algorithm.

The average success rates of this method in comparison with five clusters obtained and published by Steane *et al.* [12] are represented in Figure 1. The diagram demonstrates how the success rates depend on the choice of the local alignment metric.

3 Nearest Neighbour Clustering Algorithm with Alignment Metrics

The second algorithm we implemented is an analog of the *nearest neighbour* clustering algorithm (Witten and Frank [15], Section 4.7). The standard nearest neighbour clustering algorithm implemented in WEKA could not be applied directly to the ITS dataset, because it handles data represented as points in an n-dimensional Euclidean

space. Thus we had to encode a new version of the nearest neighbour algorithm based on optimal local alignments of the given sequences.

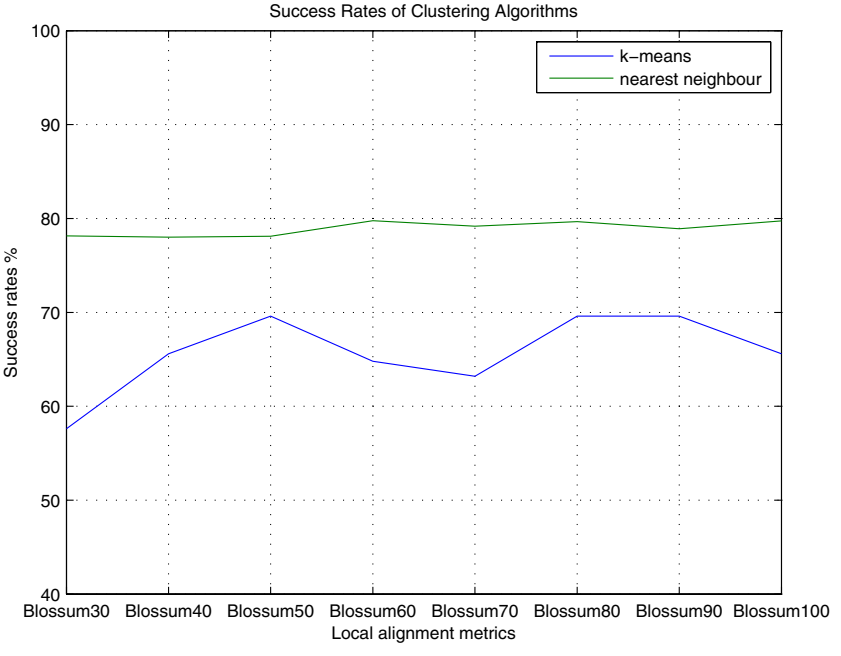

Fig. 1. Experimental data on success rates of clustering algorithms for several distance metrics

Given the number k of clusters, the algorithm computes all local alignment scores between all pairs of DNA sequences in the data set. It chooses random k sequences as representatives of the k clusters. For every other sequence s in the data set, it looks at all sequences which have been considered and allocated to the clusters and finds the nearest neighbour of s among these sequences. The algorithm then allocates s to the cluster of its nearest neighbour. This is repeated until all sequences have been assigned to clusters.

The average success rates of this method using various alignment metrics compared with the clusters obtained and published by Steane et al. [12] are represented in Figure 1.

4 Experimental Results

We investigated the groupings of an ITS dataset, displayed in Figures 2, 3 and 5 of Steane *et al.* [12]: The dataset includes many of different species from all subgenera and sections of *Eucalyptus*, as well as some other genera that are closely related to *Eucalyptus*. For a detailed description of the dataset we refer to [12].

Figure 1 summarizes the results of comparison between classifications obtained by our two algorithms and known classifications considered in the biological literature (see Steane *et al.* [12]). Our clustering algorithms use the strong similarity measures and have achieved high accuracy rates compared to other algorithms considered in other similar situations previously (see [14]).

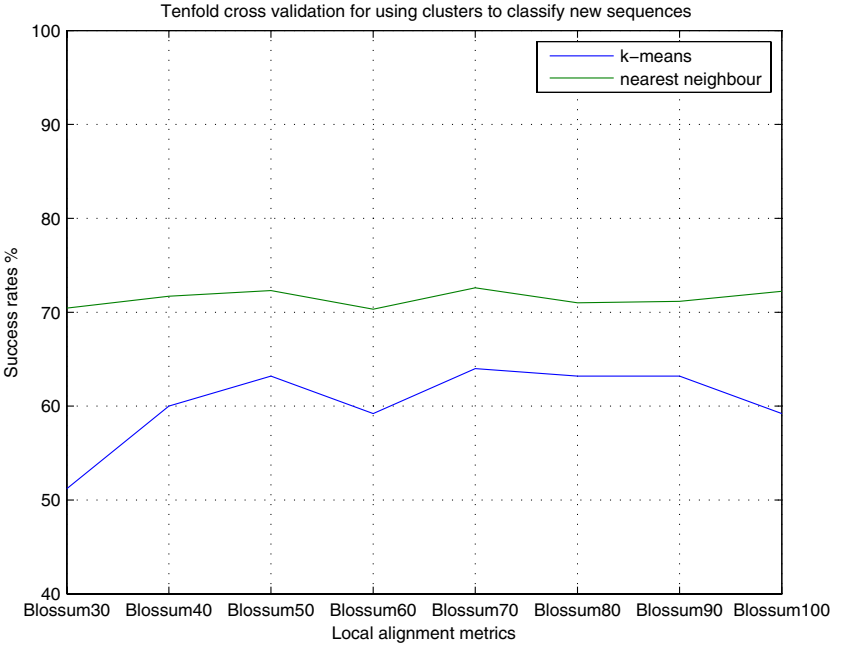

Fig. 2. Tenfold cross validation for the accuracy of allocating new sequences

We investigated the efficiency of the clusters produced by our algorithms for classifying new sequences (Figure 2), and established that both algorithms are accurate. The choice of algorithm is a trade-off between speed and accuracy. The nearest neighbor clustering algorithm is more accurate but is slower.

This research has been supported by the IRGS grant K14313 of the University of Tasmania.

References

1. Baldi, P. and Brunak, S.: Bioinformatics : The Machine Learning Approach. Cambridge, Mass, MIT Press, (2001).
2. Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G.: Biological Sequence Analysis. Cambridge University Press (1999).
3. Gedeon, T.D. and Fung, L.C.C.: AI 2003: Advances in Artificial Intelligence. Proc. 16th Australian Conference on AI, Perth, Australia, December 3-5, 2003. Lecture Notes in Artificial Intelligence 2903 (2003).

4. Gusfield, D.: Algorithms on Strings, Trees, and Sequences. Computer Science and Computational Biology, Cambridge University Press, Cambridge (1997).
5. Jones, N.C. and Pevzner, P.A.: An Introduction to Bioinformatics Algorithms. Cambridge, Mass, MIT Press, (2004). http://www.bioalgorithms.info/
6. Kang, B.H.: Pacific Knowledge Acquisition Workshop. Part of the 8th Pacific Rim Internat. Conf. on Artificial Intelligence, Auckland, New Zealand, (2004).
7. Kang, B.H., Kelarev, A.V., Sale, A.H.J. and Williams, R.N.: A model for classifying DNA code embracing neural networks and FSA. Pacific Knowledge Acquisition Workshop, PKAW2006, Guilin, China, 7-8 August 2006, 201-212 (2006).
8. Kelarev, A.V.: Ring Constructions and Applications. World Scientific, River Edge, NJ, (2002).
9. Kang, B.H., Kelarev, A.V., Sale, A.H.J. and Williams, R.N.: Labeled directed graphs and FSA as classifiers of strings. 17th Australasian Workshop on Combinatorial Algorithms, AWOCA 2006, 93-109 (2006).
10. Li, J., Yang, Q. and Tan, A.-H.: Data Mining for Biomedical Applications. Proc. PAKDD 2006, BioDM 2006, Singapore, April 9, 2006. Lecture Notes in in Artificial Intelligence 3916 (2006).
11. Mount, D.: Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory, (2001). http://www.bioinformaticsonline.org/
12. Steane, D.A., Nicolle, D., Mckinnon, G.E., Vaillancourt, R.E. and Potts, B.M.: High-level relationships among the eucalypts are resolved by ITS-sequence data. Australian Systematic Botany 15, 49-62 (2002).
13. Webb, G.I. and Yu, X.: Advances in Artificial Intelligence: Proc. 17th Australian Joint Conference on Artificial Intelligence AI 2004, Cairns, Australia, December 4-6, 2004. Lecture Notes in Artificial Intelligence 3339 (2004).
14. WEKA, Waikato Environment for Knowledge Analysis, http://www.cs.waikato.ac.nz/ml/weka, viewed 20.06.2006.
15. Witten, I.H. and Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Elsevier/Morgan Kaufman, Amsterdam (2005).
16. Yang, J. Y. and Ersoy, O.K.: Combined supervised and unsupervised learning in genomic data mining, School of Electrical and Computer Engineering, Purdue University, (2003).
17. Zhang, C., Guesgen, H.W. and Yeap, W.K.: Trends in Artificial Intelligence. 8th Pacific Rim Internat. Conf. on Artificial Intelligence PRICAI 2004, Auckland, New Zealand, August 9-13, 2004. Lecture Notes in Artificial Intelligence 3157 (2004).
18. Zhang, S. and Jarvis, R.: Advances in Artificial Intelligence. Proc. 18th Australian Joint Conference on Artificial Intelligence AI 2005, Sydney, Australia, December 5-9, 2005. Lecture Notes in Artificial Intelligence 3809 (2005).