A new asymmetric, space variant distance metric

Shobhit Saxena 2002407
Rahul Malik 2002442

January 20, 2007

1 Introduction to the problem

Distance measures play a vital role in many applications such as supervised and unsupervised learning, information retrieval, and product recommendations. Typically, the distance measures are symmetric, and do not adapt based on the point from which it is measured. However, many practical applications need the measures to be asymmetric and vary with the context. For example, the measure used to find documents similar to a document on botany should be different from the measure used in the case of a document on zoology. Similarly, the measure of inclusion of meaning of one sentence in another is not symmetric [2].

Clustering plays a major role in data mining as a tool to discover structure in data. Object clustering algorithms operate on a feature vector representation of the data and find clusters that are compact with respect to an assumed (dis)similarity measure between the data points in feature space. As a consequence, the nature of clusters identified by a clustering algorithm is highly dependent on the assumed similarity measure. The most commonly used dissimilarity measure, namely the Euclidean metric, assumes that the dissimilarity measure is isotropic and spatially invariant, and it is effective only when the clusters are roughly spherical and all of them have approximately the same size, which is rarely the case in practice [3]. The problem of finding nonspherical clusters is often addressed by utilizing a feature weighting technique. These techniques discover a single set of weights such that relevant features are given more importance than irrelevant features. However, in practice, each cluster may have a different set of relevant features.

2 Previous work

Traditionally, the Euclidean and Mahalanobis distance metrics have been used for machine learning applications. The Mahalanobis distance is based on the correlations between variables by which different patterns can be identified and analysed. It is a useful way of determining similarity of an unknown sample set to a known one. It differs from Euclidean distance in that it takes into account

the correlations of the data set and is scale-invariant, i.e. not dependent on the scale of measurements

Another class of distance metrics, context-sensitive learnable asymmetric dissimilarity (CLAD) measures [1] are defined to be a weighted sum of a fixed number of dissimilarity measures where the associated weights depend on the point from which the dissimilarity is measured. The parameters used in defining the measure capture the global relationships among the features. The dissimilarity measures can be learned automatically from a set of user specified comparisons which are in the form "x is closer to y than to z".

3 Action Plan

- First we intend to perform a literature study of the various distance measures and how they suit the various clustering algorithms. Then we intend to propose a new distance metric based on our study.
- Next we would perform the implementation work to have a platform to test our proposed metric.
- We'll finally perform testing and analysis using various data sets and determine the usefulness of our measure. To do this, we will compare our results against other standard measures.

4 Resources

We will use 20 News groups dataset for testing our algorithm. 20 News Group data contains clusters of varying degrees of separation.

Data source: 20 News groups http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.tar.gz.

References

[1] Krishna Kummamuru, Raghu Krishnapuram, Rakesh Agrawal. On Learning Asymmetric Dissimilarity Measures, *Proc. Fifth IEEE International Conference on Data Mining (ICDM05)* .

[2] K. Krishna and R. Krishnapuram. A clustering algorithm for asymmetrically related data with applications to text mining, *In Proceedings of CIKM, pages 571573. ACM, 2001.*

[3] R. Krishnapuram and J. Kim. A note on fuzzy clustering algorithms for Gaussian clusters. *IEEE Transactions on Fuzzy Systems, 7(4):453461, Aug 1999.*