A New Homogeneity Inter-Clusters Measure in Semi-Supervised Clustering

Abstract

Many studies in data mining have proposed a new learning called semi-Supervised. Such type of learning combines unlabeled and labeled data which are hard to obtain. However, in unsupervised methods, the only unlabeled data are used. The problem of significance and the effectiveness of semi-supervised clustering results is becoming of main importance. This paper pursues the thesis that much greater accuracy can be achieved in such clustering by improving the similarity computing. Hence, we introduce a new approach of semi-supervised clustering using an innovative new homogeneity measure of generated clusters. Our experimental results demonstrate significantly improved accuracy as a result.

References

- Jain, A. K. and Dubes, R. C.: Algorithms for clustering data, Prentice-Hall, Inc., Upper Saddle River, NJ, USA. (1988).
- Zhang, Y., Mao, J., and Xiong, Z. 2003: An Efficient Clustering Algorithm. International Conference on Machine Learning and Cybernetics, vol. 1, pp. 261-265.
- Hartigan, J. A. 1975. Clustering Algorithms. Wiley, New York.
A New Homogeneity Inter-Clusters Measure in Semi-Supervised Clustering

- Anderberg, M. R. 1973. Cluster Analysis for Application. Academic Press, New York.
- KiriWagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl, “Constrained K-means Clustering with Background Knowledge”; ICML ’01 Proceedings of the Eighteenth International Conference on Machine Learning, 2001
- SugatoBasu, ArindamBanerjee, and Raymond J. Mooney. Semisupervisedclustering by seeding. In Proceedings of 19th InternationalConference on Machine Learning, 27-34.
- Klein, D. , S. D. Kamvar, and C. D. manning: From instance-levelconstraints to space-levelconstraints: making the most of priorknowledge in dataclustering. In the international conference on machine learning, (ICML ’02), Springer-Verlag. San francisco, USA (2002) 307-314.
- Han, J. , and Kamber, M. 2000. Data Mining: Concepts and Techniques. Morgan Kaufmann.
- Bilenko, M. , Basu, S. , and Mooney, R. J. (2004). Integrating constraints,and metric learning in semi-supervised clustering. In ICML ’04: Proceedings of the twenty-first international conference on Machine learning, pages 81–88, New York, NY, USA. ACM.
- (Kestler et al, 2006)Kestler, H. A. , Kraus, J. M. , Palm, G. and Schwenker, F. : On the e?ects of constraints in semi-supervised hierarchical clustering. Arti?cial Neural Networks in Pattern Recognition, Springer-Verlag. Germany (2006) 57-66
- Bade, K. , Hermkes, M. and Nrnberger. A: User orientedhierarchicalinformationorganization and retrieval. Proceedings of the European conference on Machine Learning, (ECML’07), Springer-Verlag. Poland (2007) 518-526.
- Daniels, K. and Giraud-Carrier, C. (2006). Learning the threshold in hierarchical agglomerative clustering. In ICMLA ’06: Proceedings of the 5th International Conference on Machine Learning and Applications, pages 270–278, Washington, DC, USA. IEEE Computer Society.
- Bohm, C. and Plant, C. : Hissclu : A hierarchicaldensity-basedmethod for semi-supervised clustering. Proceedings of the international conference on extendingdatabasetechnology, (EDBT ’08). New York, USA (2008) 440-451.
- Davidson I., Ravi, S. S., Intractability and Clustering with Constraints. To Appear in the Proceeding of ICML 2007.
- MP. Murphy and D. W. Aha, Ucirepositorydatabases. http://www. ics. uci. edu/mlearn, 1996.
- Quinlan, J. R. : Induction of decisiontrees. Machine Learning (1986) 81-106.

Index Terms

Computer Science
Artificial Intelligence
Keywords
Semi-supervised clustering distance computation homogeneity measure