COMP6036: Advanced Machine Learning

An investigation into DBSCAN

Ashley J. Robinson
ajr2g10@ecs.soton.ac.uk

School of Electronics and Computer Science
University of Southampton

3rd March, 2014

Abstract

Your abstract goes here

1 Motivation for Algorithm

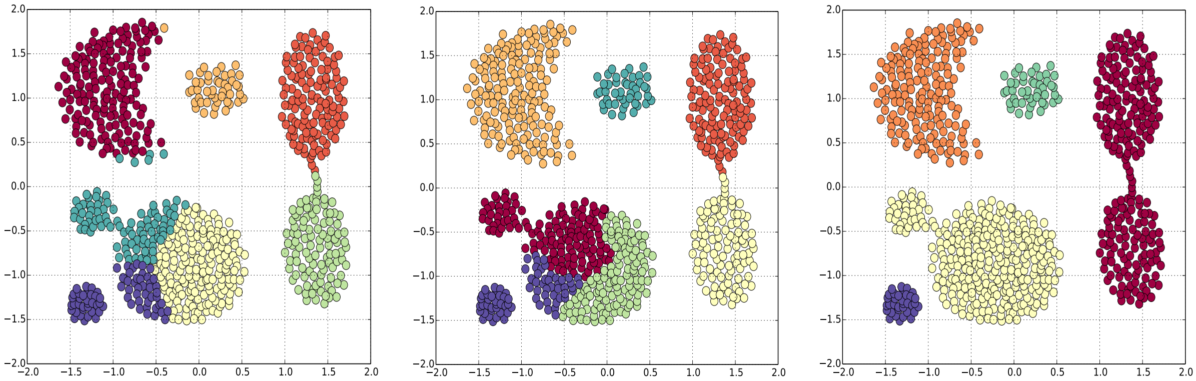
DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is an application of machine learning introduced by Ester et al. (1996). Intended to address Spatial Database Systems (SDBS) which can be produced from natural geometric and geographical datasets or applications such a layout for integrated circuit design (Güting, 1994). It has three main objectives. To minimise the required domain knowledge needed to set input parameters, have the capability to discover clusters of arbitrary shapes and to perform well on large spatial databases.

At the time of creation the algorithms was compared to a recent development called CALARANS (Raymond and Jiawei, 1994) which is an extension of CLARA (Clustering LARge Applications) (Kaufman and Rousseeuw, 1990). Both algorithms are intended for use on large databases but CLARANS uses random noise to improve performance. Apart from traditional clustering algorithms, such a K-means, DBSCAN was a breakthrough in terms of a density approach to datasets.

2 Technical Explanation

DBSCAN uses cluster density to classify data. It tries to find connectivity between data poiunts within a reasonable distance. This distance is passed to algorithm as an input parameter and can be tuned to give

The graphs in Figure 1 are different clustering algorithms applied to a shape dataset taken from Gionis et al. (2005). Algorithms were taken from a machine learning toolbox for Python (scikit learn, 2013).



(a) K-Means (b) WARD (c) DBSCAN

FIGURE 1: A comparison against DBSCAN

The algorithm exhibits variation in perforamce as per the standard bias-variance dilemma. Figure 2 shows how the error varies for DBSCAN when applied to data used in Figure 1. The EPS paramter sets the distance which the algorithm allow itself to jump and consider itself in the same cluster.

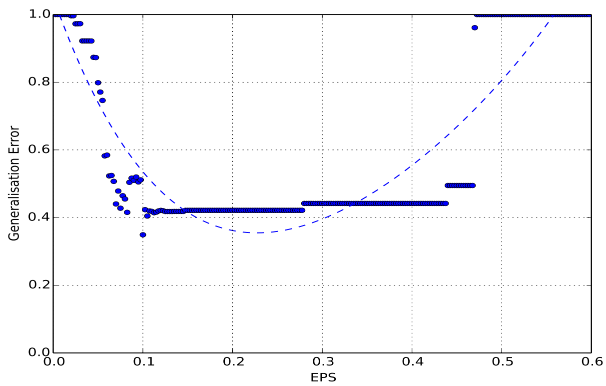


FIGURE 2: Generalisation error with varying input parameter

3 Conclusion and Further Work

References

Martin Ester, Hans peter Kriegel, Jrg S, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231. AAAI Press, 1996.

Aristides Gionis, Heikki Mannila, and Panayiotis Tsaparas. Clustering aggregation. In *In Proceedings of the 21st International Conference on Data Engineering (ICDE)*, pages 341–352, 2005. Hosted by Speech and Image Processing Unit, University of Eastern Finland, `http://cs.joensuu.fi/sipu/datasets/` (Visited on 27/02/2014).

Ralf Hartmut Güting. An introduction to spatial database systems. *Very Large Database Journal (VLDB J.)*, 3(4):357–399, 1994.

Leonard Kaufman and Peter J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis.* John Wiley and Sons, New York, 1990.

Ng T. Raymond and Han Jiawei. Efficient and effective clustering methods for spatial data mining. In *20th Very Large Database Conference, Santiago, Chile, 1994*, pages 144–155, 1994.

scikit learn. machine learning in python. `http://scikit-learn.org/stable/index.html`, 2013. (Visited on 02/27/2014).