Testing and Tuning SkinnyDip: Noise-Robust Clustering

Grant King
kinggra1@msu.edu

1. PROBLEM DESCRIPTION

It is not uncommon for some real-world data sets to feature an abundance of noise. Depending on the severity of this noise, classical clustering methods may fail due to due to dependence on clean data or confusion from increasingly excessive noise. SkinnyDip is an algorithm proposed by Maurus et al.[**?**] designed to handle clustering data in highly noisy environments. SkinnyDip is based on the statistical concept of the *dip* [2].

The dip views the structure of the Empirical Cumulative Distribution Function(ECDF) of a set of single dimensional data to determine whether it is unimodal or multimodal. This test is expanded to a multidimensional, recursive heuristic in order to isolate the various modes of each feature of a set of data. This results in a deterministic, parameter-free, unsupervised method of finding clusters that are based on the modes of multivariate distributions.

A visual example of the purpose of SkinnyDip is demonstrated in the running example from [**?**] which shows the extraction of distinct shapes from a two-dimensional data field that consists of 80% static noise. A variety of similar clustering methods were shown to perform poorly on this data set, both through the inclusion of the evenly-distributed, static noise in clusters, and through over-segmentation of the existing classes.

My goal is to implement the SkinnyDip algorithm and perform additional tests to demonstrate its usefulness on noisy, real-world data sets, with a final goal of reducing the extraneous information introduced by SkinnyDip's hypercubic boundaries through extension with a second clustering method.

2. RELATED WORK

Data clustering is a fundamental problem in Machine Learning, and has been approached in a variety of ways, resulting in a plethora of tools and methods [**?**], many of which are sensitive to noise and other outliers. Additionally, many common techniques, such as k-means clustering, operate as closed set clustering methods and are unable to reject noise at all.

[**?**] is a density-based technique for finding clusters in environments that may contain noise. However, it has the disadvantages of being a parameterized method, and still continues to find extraneous clusters in increasingly noisy data when compared to SkinnyDip.

A single other existing method for clustering using the statistical dip test was found, a technique called DipMeans[**?**]. This method, however, takes a different approach to the test, and performs the dip test on a collection of distance measurements as opposed to the raw data values themselves. SkinnyDip, however, requires no distance measurements and is both functionally and computationally distinct from the DipMeans technique.

3. PROJECT MILESTONES

Projected Project Milestones are given in 2. It appears that most of the work will be in creating an actual implementation of the SkinnyDip algorithm itself in MATLAB. This may be mitigated by using an existing code base implemented in R and focusing more on the testing and extension vs. the algorithm itself, though this would require the extra learning curve of an entirely new language and tool set. Beyond this, the work that will largely be reported for this project will involve working with some real-world data (currently considering star data and galaxy clustering) as well as testing methods to improve the accuracy of SkinnyDip under higher-dimensional data sets.

4. REFERENCES

[1] Seunghoon Hong, Tackgeun You, Suha Kwak, and Bohyung Han. Online tracking by learning discriminative saliency map with convolutional neural network. *arXiv preprint arXiv:1502.06796*, 2015.

[2] J. A. Hartigan and P. Hartigan. The dip test of unimodality. *The Annals of Statistics*, 1985.

[3] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[4] Jialue Fan, Wei Xu, Ying Wu, and Yihong Gong. Human tracking using convolutional neural networks. *Neural Networks, IEEE Transactions on*, 21(10):1610–1623, 2010.

[5] Naiyan Wang and Dit-Yan Yeung. Learning a deep compact image representation for visual tracking. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 809–817. Curran Associates, Inc., 2013.

[6] H. Li, Y. Li, and F. Porikli. Deeptrack: Learning discriminative feature representations online for robust visual tracking. *Image Processing, IEEE Transactions on*, PP(99):1–1, 2015.

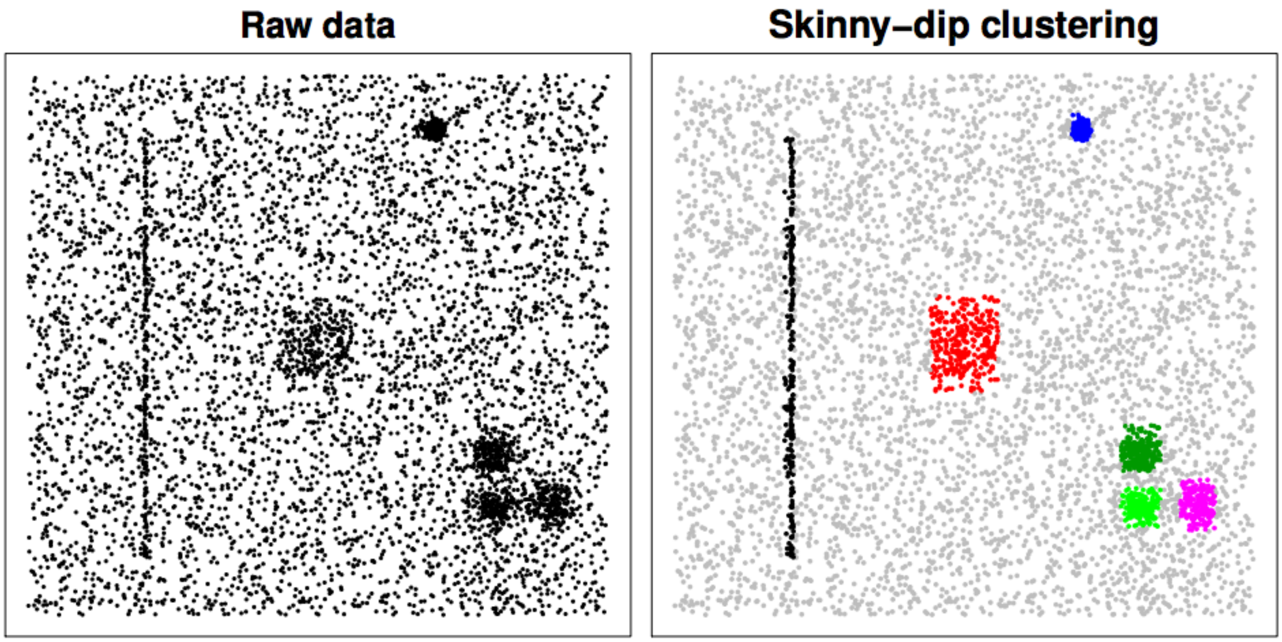

Figure 1: The running example used by Maurus, et al. throughout [?]

2/17/2017	Proposal Due
2/24/2017	Finalize Project Goals
3/3/2017	Working Code for UniDip
3/10/2017	Testing on Synthetic Data with UniDip
3/17/2017	Intermediate Report Due
3/24/2017	Find Noisy Real-World Data Sets
3/31/2017	Expand to SkinnyDip
4/7/2017	Expand to SkinnyDip
4/14/2017	Try Fine Tuning Clustering Results (e.g. Density Based Methods)
4/21/2017	Finalize Report

Figure 2: Tentative Project Milestones