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.[1] 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.

My goal is to augment the SkinnyDip algorithm with the addition of a Gaussian clustering model in an attempt to reduce the excessive inclusion of noise in hypercubic-bound clusters, and perform additional tests to demonstrate its usefulness on noisy, real-world data sets.

2. INTRODUCTION

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 [3], 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.

[4] 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[5]. This method takes a different approach to using the dip test by performing it 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.

While SkinnyDip does manage to find all values within a cluster with high accuracy and precision, there is still a notable risk of falsely matching noise to a particular class. This is inherent in the way that the algorithm segments the clusters into hypercubic regions. If the data does not fit a cubic model, then the values included in extreme regions of the cluster (e.g. the corners of a square cluster) are likely to be false positives and should not have been included. This discrepancy increases exponentially as the dimension of the data increases, an issue acknowledged in the original paper[1]. I plan to demonstrate that the addition of a second step to clustering can reduce this particular error rate, with minimal negative impact on the existing true positive matches.

3. DATA

A visual example of the purpose of SkinnyDip is demonstrated in the running example data from [1] (See Figure 1) 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 excessive segmentation of the actual classes.

Aside from the uniform background noise, the running example data consists of 2 distinct cluster models: two rectangular model clusters and four 2D Gaussian model clusters. In Figure 1 the square classes are colored red and black, and the remaining four, smaller clusters are the 2D Gaussians.

4. PROJECT MILESTONES

Projected Project Milestones are given in Figure 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.

In order to obtain a solid understanding of the algorithm itself, I am inclined to perform a migration of the code, which may also help determine where future extensions should be worked in. For an additional level of understanding, I plan on spending some time before the intermediate milestone testing the dip-based clustering method on synthetic data sets to determine what kinds of data may cause issues.

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.

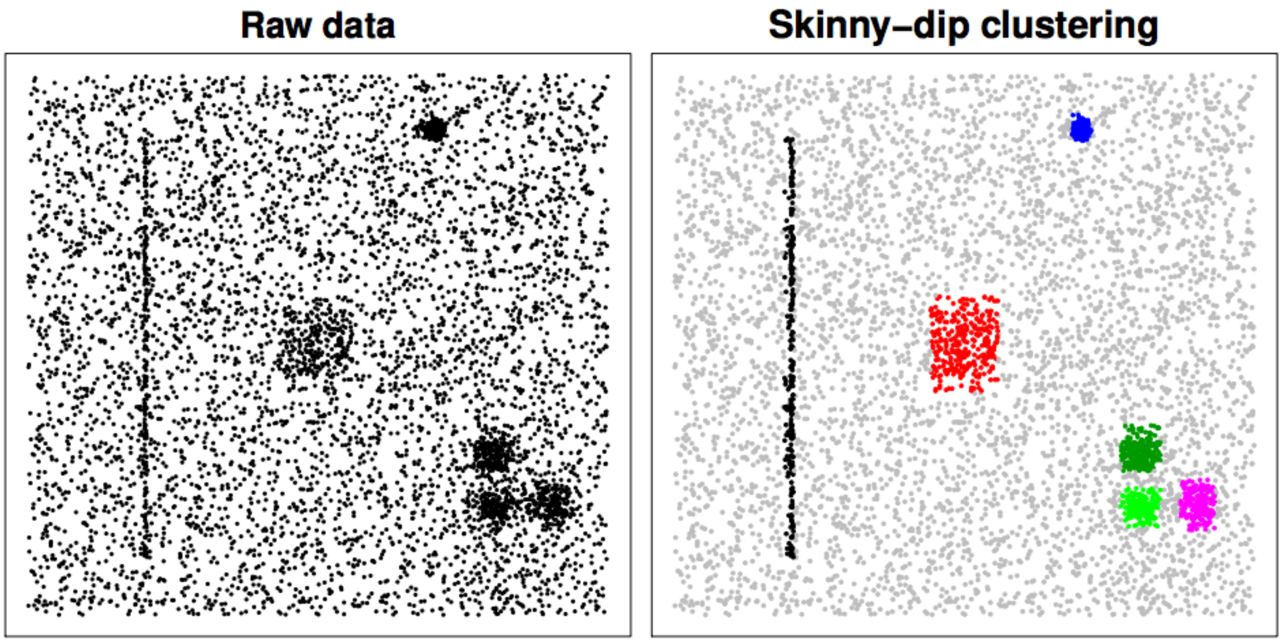

Figure 1: The running example used by Maurus, et al. throughout [1]

5. REFERENCES

[1] Samuel Maurus and Claudia Plant. Skinny-dip: Clustering in a sea of noise. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1055–1064, 2016.

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

[3] Rui Xu and D. Wunsch, II. Survey of clustering algorithms. *Trans. Neur. Netw.*, 16(3):645–678, May 2005.

[4] Martin Ester, Hans peter Kriegel, JÃűrg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231, 1996.

[5] A. Kalogeratos and A. Likas. Dip-means: an incremental clustering method for estimating the number of clusters. *In Advances in neural information processing systems, pages 2393âĂŞ2401*, 2012.

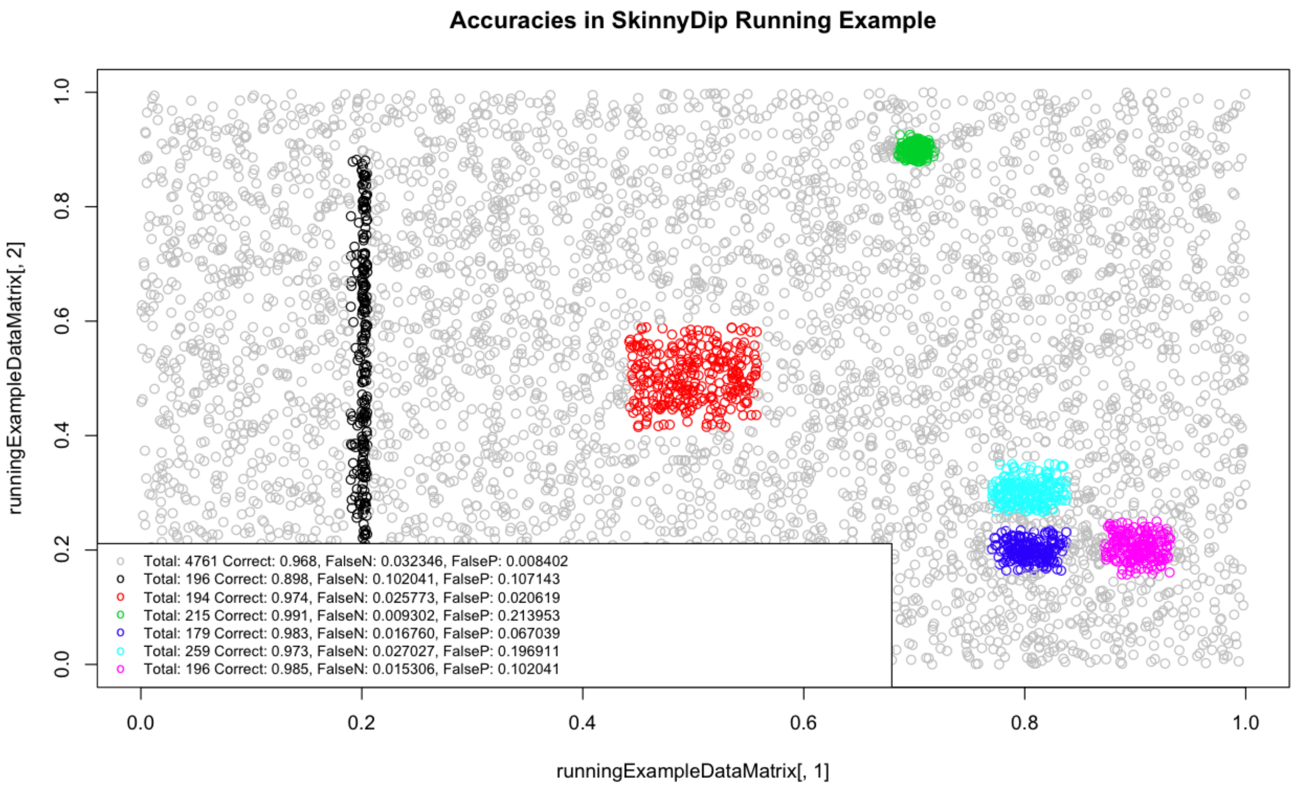

Figure 2: Accuracies of the given SkinnyDip example

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 3: Tentative Project Milestones