

Density Biased Sampling

An Improved Method for Data Mining
Clustering

Authors

Christopher R. Palmer and Christos Faloutsos

Presenters

Renee Szwedo and Axel Arditi

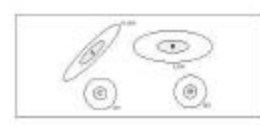
Introduction – The Problem

- Clustering is an excellent way to Mine Data for interesting results
- Clustering is difficult and inefficient
- Sampling has been around in Statistics for a long time, and is a common way to decrease the size of a Dataset.
- Reducing the size of the Dataset eases the burden of Clustering

Introduction – More Problems

- Sampling is oftentimes not a good enough representation of the original Data; too much is left out.
- Uniform Sampling (Samples of data points chosen with equal probability) is effective only if the data itself is evenly distributed as well.

Uniform Sampling – Not Enough



If the above dataset is sampled using Uniform Sampling techniques, clusters C and D will most likely be dismissed as background noise.

So What?

- The large clusters are oftentimes information we already know. Simple data is easy to segregate, but with more complex data types (20+ dimensions) it is near impossible to weed out these large clusters.
- Sometimes, those small clusters contain the most interesting information.

Key Contribution

- Uniform Sampling performs well when Data is already grouped into even Clusters.
- Uniform Sampling does NOT perform well when the Data is grouped into skewed Clusters.
- Contribution presents a new Sampling technique that does better than Uniform Sampling Methods on skewed Datasets.

Prior Work

- Uniform Sampling and Clustering has been thoroughly researched.
- Most recent and successful, “BIRCH” algorithm, proposed by Zhang, Ramakrishnan and Miron involves a data summarization step before Clustering
- Density Biased Sampling compared directly against BIRCH in Results Section.

More Prior Work

- Density Biased Sampling (DBS) directly related to Probability Proportional to Size (PPS) Sampling.
- PPS Sampling is a multi-stage sampling technique involving many passes through the Data.
- DBS, inversely biased by group size, is a one-stage Sampling Technique

Methodology

- Goal – Summarize the data for easy Clustering
- Data must be pre-partitioned into “groups”
- See Generalization Rules taught earlier in the semester
- Once data is generalized, a sample may be taken.

Methodology

- Random Sampling assumes each point has equal probability of being chosen in the sample.
- Here, we come up with a probability function inversely dependant on the size of the “group.”
- $f(n) = P(x)$, x in Group N of size n

Methodology

- With a probability function dependant on group size, each point within a group has the same probability of being chosen.
- Hence, within a group, all points are selected uniformly.
- Points are then assigned weights based on their group size.

Methodology

- All points within a group have equal weight; this preserves the original density
- The proposed weighting function is the inverse of the probability function, $f(x)$.
- Weight $w(x) = 1/f(x)$

Methodology

- To determine the probability that any given point will be brought into the Sample, multiply its weight with its original Probability. Since weight and probability are uniform amongst points within a group, the sum of all probabilities should equal the group size.

$$\sum_{j=1}^{n_i} P(\text{point } x_j) \cdot w(n_i) = \sum_{j=1}^{n_i} f(n_i) \cdot 1/f(n_i) = n_i$$

Methodology

- To allow the previous equation to be biased towards group size, we must define $f(x)$ accordingly:
 - $f(x) = a / n^e$
 - Where e is a constant and a reflects total Sample Size.

Methodology

- Notice that if $e=0$, this is Uniform Sampling
- a is defined as below where M is the size of the Sample

$$E(\text{sample size}) = \sum_{i=1}^g E(\text{size of group } i)$$

$$M = \sum_{i=1}^g n_i f(n_i) = \sum_{i=1}^g n_i \frac{a}{n_i^e}$$

$$\Rightarrow a = \frac{M}{\sum_{i=1}^g n_i^{1-e}}$$

Results

- The results published were “good.”
- The aim was to prove that this technique out-performs uniform sampling on datasets of non-uniform partitioning.
- Graphs show three distributions; DBS outperforms Uniform Sampling in two of three scenarios.

Results – Graph Explanation

- The X-axis represents increasing Sample Size as a percentage. The percentage indicates the percent of the total points the sample is to be.
- The Y-axis represents the number of centroids found within 0.001 distance of known correct centroids within the data.

Opinion

- The paper was concise, but at some times, too concise. This is to be expected considering the intended audience.
- Overall, we rate this paper 7 of 10; the topics were undeniably interesting but the presentation was oftentimes too complicated and unclear.