Density-Biased Clustering Based On Reservoir Sampling

Kittisak Kerdprasop, Nittaya Kerdprasop and Pairote Sattayatham
Data Engineering and Knowledge Discovery (DEKD) Research Unit[*]
School of Computer Engineering, Suranaree University of Technology, Thailand
{kerdpras, nittaya, pairote}@ccs.sut.ac.th

Abstract

Clustering is a task of grouping data based on similarity. A popular k-means algorithm groups data by firstly assigning all data points to the closest clusters, then determining the cluster means. The algorithm repeats these two steps until it has converged. We propose a variation called weighted k-means to improve the clustering scalability. To speed up the clustering process, we develop the reservoir-biased sampling as an efficient data reduction technique since it performs a single scan over a data set. Our algorithm has been designed to group data of mixture models. We present an experimental evaluation of the proposed method.

1. Introduction

Clustering is the automatic grouping of data based on similarity. There exists a large number of clustering techniques, but the most classical and popular one is the k-means algorithm [4]. Given a data set containing n objects, k-means partitions these objects into k groups. Each group is represented by the centroid of the cluster. Once cluster representatives are selected, data objects are assigned to the nearest centers. The algorithm iteratively selects new better representatives and reassigns data objects until no change is made. At this point the algorithm is said to converge. Even though k-means is an effective clustering algorithm, it can sometimes converge to a local optimum. Many methods [2, 12] have been developed to extend the k-means with the common objective of avoiding converging to a bad local optimum. Some methods [5, 7] search for the best initialization because k-means is known to be sensitive to initial point selection.

Another difficulty of clustering with k-means is that it fails to identify clusters with large variation in sizes since original large clusters tend to be split. Clustering algorithms, such as DBSCAN [8] and CURE [1], have been developed to overcome this kind of difficulty. However, with very large data set, these algorithms degrade considerably.

When clustering massive data set, data reduction is an effective technique to speed up the algorithm. Sampling is a powerful data reduction paradigm to remedy the inherent complexity of clustering. Uniform random sampling in which every data point has the same probability of being selected has been used extensively in data mining and databases [9, 10]. In the case of data sets with large variation in cluster sizes, density biased sampling [3, 6] tends to be a better scheme. In density biased sampling, the probability that a data point will be included in the sample is varied by the density of a cluster. Our work also follows this path with a step further on extending the k-means algorithm to work with a weighted sample. We propose an algorithm on density biased sampling based on the reservoir technique and a weighted k-means algorithm to cluster a data sample augmented with weights.

2. Data reduction

On scalable popular and successful clustering methods such as k-means to work against large data sets, many algorithms employ the sampling technique to minimize data sets. The sampling technique used in these algorithms is uniform random sampling, which assigns every object the same probability of being included in the sample. But many data sets in real life do not follow the uniform distribution scheme. It instead seems to follow the Zipf's distribution [13], for instance, income and population distribution. In these data sets, some areas such as large metropolitan area have much higher population density than the small

[*] DEKD Research Unit is fully supported by the research grants from Suranaree University of Technology. This work has been supported by grants from the Thailand Research Fund (TRF, MRG4780170), and the National Research Council.

cities. If all the populations have equal opportunity of being selected as a representative, sparse areas may be missed and not be included in the sample.

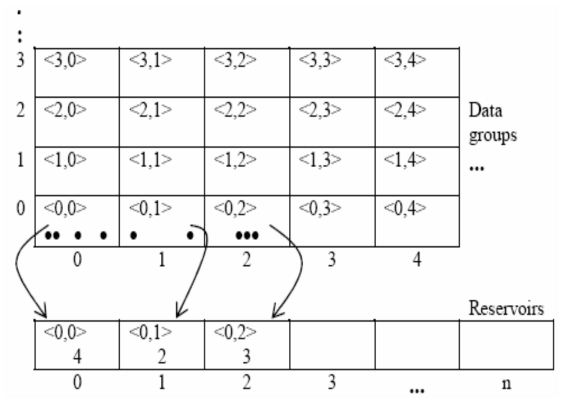
Density biased sampling [6] is a sampling technique that takes into account the different sizes of the groups. Small groups or sparse regions are assigned higher probability to be included in the sample than the large groups or dense regions. By biasing the sampling process, small clusters will not be missed or overlooked as outliers. Palmer and Faloutsos [6] develop a non-uniform sampling method for clusters that differ very much in size and density. Their method is a generalization of uniform random sampling in that every group of data sets can be assigned different probability of being drawn. When sampling is biased by group density, smaller groups are over-sampling, whereas larger groups are under-sampling. However, their method is significantly affected by noise due to the tendency of over-sampling noisy area.

We propose a novel approach of adapting reservoir technique [11] to perform a density biased sampling on large data sets. Our algorithm can obtain a desired sample through a single data set scan. The proposed method is simpler and requires less resource than the hash-based method [6].

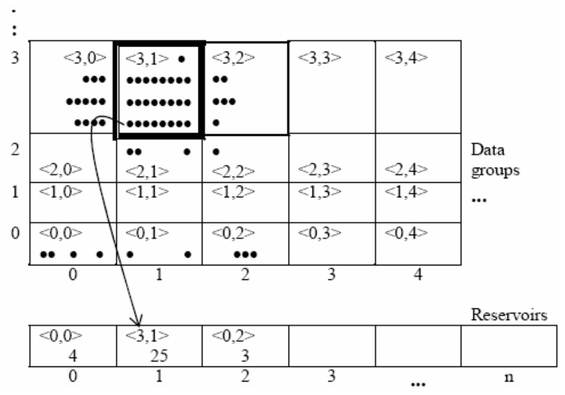
A reservoir-sampling algorithm is a simple, unbiased random sampling algorithm for drawing a sample of size n without replacement from a population of size N ($N \geq n$). Vitter [11] has developed a one-pass reservoir-sampling algorithm when the population size (N) is unknown and cannot be determined efficiently. The term "reservoir" defines a storage area j ($j \geq n$, but mostly $j = n$) to store the potential candidates of the sample. The j reservoirs is initialized to store the first j records of the file, that is, all areas of the reservoir pool are initially filled up. Then the algorithm starts scanning the remaining part of the file with a randomly skipping step. The random picked record is evaluated whether to replace the existing one in the reservoir pool. If it passes the test, the position in the reservoir is also randomly selected. The process stops when the end of file has been reached and the records in the reservoir form a simple random sample of the population.

Our sampling algorithm generalizes the reservoir scheme for the case of data with different density distribution. In our proposed method, the initial step of partitioning data into groups resembles that of Palmer and Faloutsos [6]. But our subsequent steps are not based on hashing scheme in order to avoid the effect of noise and collision problems.

After the initial step of dividing the data space into bins of equal size, the information of the first n groups are put into the n reservoirs residing in main memory (see Figure 1a). The collected information includes the number of points in each group and the id of the group.



(a) initialize the reservoir



(b) update reservoir randomly

Figure 1. Density biasing in a reservoir scheme

The algorithm performs a single scan on a data set in a random manner controlled by a random variate S with the distribution W. The density biasing (step 7 in Figure 2) is achieved through the consideration of two consecutive data groups. If the density difference of the two data groups is above some threshold δ (i.e., detecting cluster edge) or the sum of density on both groups is above the threshold value ε (i.e., avoiding noisy cases), then the denser group is a candidate to be included in a sample. This new candidate is put into a reservoir pool at a random position (the reservoir update is pictorially shown in Figure 1b). The density-biased sampling proceeds until the skipping variate S reaches the end of the data groups.

```
Algorithm Density-biased reservoir sampling
    Input:   a data set of $N$ objects
    Output:  a density-biased sample of size $n$ ($n \le N$)
             associated with weight $w$
1) Partition data into $g$ groups (with group-id $1,2,...,g$), $g \ge n$
2) Initialize the reservoir $X_1, ..., X_n$ to be the first $n$
   <group-id, density>-pairs of the data groups
3) Set $W \leftarrow \exp(\log(\text{random}()) / n)$   // initialize $W$ that
   // will be used in the generation step of random variate $S$
4) Set $S \leftarrow \lfloor \log(\text{random}()) / \log(1-W) \rfloor$
5) While $S < g$ do
6)     Read data groups $g_S$ and $g_{S+1}$
           // read two consecutive data groups
7)     If $(\|\text{density}(g_S) - \text{density}(g_{S+1})\| > \delta)$ OR
           $((\text{density}(g_S) + \text{density}(g_{S+1})) > \varepsilon)$
           // $\delta$ and $\varepsilon$ are predefined density threshold values
       Then $X_{1+\lfloor n*\text{random}()\rfloor} \leftarrow$ <group-id, density> of
           maximum density$\{g_S, g_{S+1}\}$
           // randomized the reservoir area to be updated
8)     $W \leftarrow W * \exp(\log(\text{random}()) / n)$
           // update $W$ for the skipping process
9)     $S \leftarrow \lfloor \log(\text{random}()) / \log(1-W) \rfloor$ // generate $S$ to
           // denote the number of groups to be skipped over
10) Return $X_1, ..., X_n$
```

Figure 2. Density-biased reservoir sampling algorithm

3. Density-biased clustering

The classical k-means algorithm [4] is a fast method to perform clustering. The original n data points to be clustered are contained in the dataset $X = \{x_1, ..., x_n\}$. The k-means algorithm partitions n data points into K sets. The assignment of a data point x_i to its nearest cluster center c_j is decided on the basis of the membership function, $m(c_j|x_i)$. The function returns either one of the {0,1} values: $m(c_j|x_i) = 1$ if $j = argmin_k||x_i - c_k||^2$; it is zero, otherwise. The new centroids of clusters can be computed from all data points x_i in the cluster. The objective function J of the algorithm is to minimize the sum of error squared, $J = \sum_{i=1}^{n} \min_{j\in\{1..k\}} \| x_i - c_j \|^2$.

In k-means algorithm, every data point has equal importance in locating the centroid of the cluster. This property does no longer hold in the case of density-biased sample clustering, for which each data point represents varied density in the original data. Therefore, the clustering algorithm has to consider a weight associated with each data point in the computation of cluster centers. The proposed extension to the k-means algorithm is called weighted k-means. Figure 3 outlines the algorithm. The membership function in the weighted k-means resembles that of the k-harmonic means algorithm [12]. The weight function in our algorithm is introduced, however, for the different purpose. It represents the density of the original data points.

```
Algorithm Weighted k-means
    Input:  a set of $n$ data points obtained from the
            density-biased reservoir sampling,
            and the number of clusters ($K$)
    Output: centroids of the $K$ clusters
1) Initialize the $K$ cluster centers
2) Repeat
     Assign each data point to its nearest cluster center
       according to the membership function,
       $m(c_j \mid x_i) = \frac{\| x_i - c_j \|^{-p-2}}{\sum_{j=1}^{k} \| x_i - c_j \|^{-p-2}}$
3)   For each center $c_j$, recompute the cluster center $c_j$
       using current cluster memberships and weights,
       $c_j = \frac{\sum_{i=1}^{n} m(c_j \mid x_i) w(x_i) x_i}{\sum_{i=1}^{n} m(c_j \mid x_i) w(x_i)}$
       where $w(x_i)$ is a weight associated with each
       data point
4) Until there is no reassignment of data points to new
     cluster centers
```

Figure 3. Weighted k-means algorithm

4. Experiments and results

We evaluate the performance of the proposed reservoir-based density bias sampling method against the hash-based sampling method [6]. The efficiency regarding memory usage of our reservoir-based sampling method is obviously better than the hash-based method. In the hashing scheme, some amount of memory is needed to store the hashing table in addition to the memory required for storing the drawn sample. Thus, it requires twice the amount of memory comparative to those required by our method.

Effectiveness of the proposed sampling method is examined by measuring the quality of a sample with respect to the number of correctly found clusters. The measurement *Number of Clusters found* (NC) [6] is calculated by comparing the distances of the cluster centers found by the clustering algorithm with the true cluster centers. We say that the cluster is found if the calculated distance is less than a predefined threshold (e.g., 0.001). We run clustering using the k-means algorithm. We use a synthetic data generator to generate d-dimensional data sets having *k* clusters and *N* data points. We vary d from 2 to 5, *k* from 2 to 10, and *N* from 5,000 to 100,000.

The results in Figure 4 show the NC when run clustering on various sample sizes with the presence of noise. The reported results are observed from the experiments using 3-dimensional data set having 7 clusters. One cluster contains 50,000 points and the other six clusters contain 500 points. The results obtained from other experiments on data sets with different dimensions, varied numbers of clusters and data points also conform to the one presented in Figure 4. The experimental results reveal the efficiency of the biased reservoir method especially in the presence of noise.

We evaluate the quality of the weighted k-means algorithm against the k-means by using the squared objective function. Lower value of a squared objective function reflects a better quality on clustering. The initialization step randomly selects data points as initial cluster centroids. We also consider running time of both algorithms.

The performance evaluation as shown on top of Figure 5 is obtained from running k-means and weighted k-means algorithms on 3-dimensional data sets of sizes varied from 5000, 10000, 20000, 35000, 55000, 75000, to 100000 data points. The number of clusters is set to be 10. Since all data points are used in weighted k-means algorithm, the weight function is set to be 1. The parameter p in the membership function is set to be 1.3.

The comparison on clustering quality and running time shown at the bottom of Figure 5 reveals the efficiency of running weighted k-means on density-biased sample. The experiments are performed on 10% sample of data with two methods of sampling: simple random sampling (RS) and density-biased reservoir sampling (DBS). The weight function of the weighted k-means algorithm is varied according to the density of the original data.

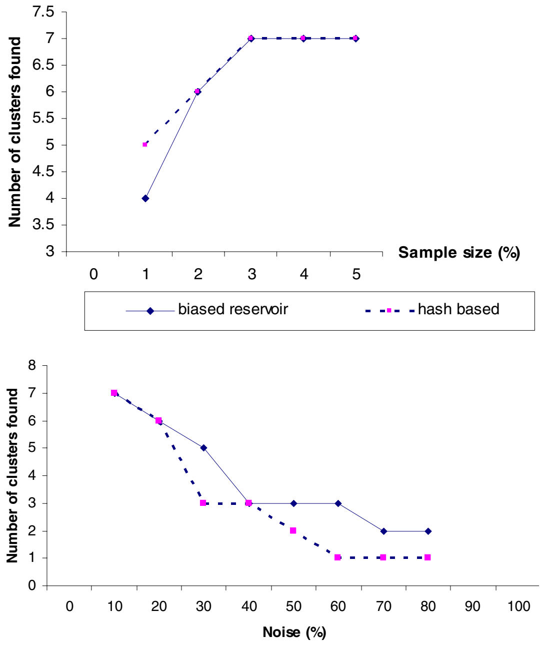


Figure 4. Finding clusters of 3-dimensional data on various sample sizes, in the presence of noise

5. Conclusions

The k-means is the simplest and most commonly used clustering algorithm. The simplicity is due to the use of squared error as the stopping criteria, which tends to work well with isolated and compact clusters. Its time complexity depends on the number of data points to be clustered and the number of iteration. We propose a variation of the k-means to better work with a large data set having much difference in cluster density. Our intuition idea is that to cope with massive data set, sampling should be the efficient data reduction method. Since the original data is assumed to be much varied in cluster sizes, density-biased sampling is an appropriate method to preserve the density.

We propose a density biased sampling technique based on the reservoir method. The inherent advantage of efficient memory usage in the reservoir scheme is adopted and extended with the additional capability of dealing with data that are much different in density distribution. The proposed technique is designed to lessen the effect of noise as it is the case in the hash-based approach. The experimental results have shown that the proposed method is as good as the hash-based method in discovering correct number of clusters. Our method, moreover, is less sensitive to noisy data.

We also develop the weighted k-means algorithm to better cluster a sample data biased by its density. The results demonstrate the efficiency of the algorithm. The evaluation of the proposed methods on real large databases and the consideration of outliers are our future work.

References

[1] S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases", *Proceedings of the ACM SIGMOD Int. Conf. on Management of Data*, 1998, pp. 73-84.

[2] G. Hamerly and C. Elkan, "Alternatives to the k-means algorithm that find better clusterings", *Proceedings of 11th ACM CIKM Int. Conf. on Information and Knowledge Management*, 2002, pp. 600–607.

[3] G. Kollios, D. Gunopulos, N. Koudas, and S. Berchtold, "Efficient biased sampling for approximate clustering and outlier detection in large data sets", *IEEE Transactions on Knowledge and Data Engineering*, 15, 2003, pp. 1-18.

[4] J. MacQueen, "Some methods for classification and analysis of multivariate observations", *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281-297.

[5] M. Meila and D. Heckerman, "An experimental comparison of model-based clustering methods", *Machine Learning*, 42, 2001, pp. 9-29.

[6] C. Palmer and C. Faloutsos, "Density biased sampling: An improved method for data mining and clustering", *Proceedings of ACM SIGMOD Int. Conf. on Management of Data*, 2000, pp. 82-92.

[7] J. Pena, J. Lozano, and P. Larranaga, "An empirical comparison of four initialization methods for the k-means algorithm", *Pattern Recognition Letters*, 20, 1999, pp. 1027-1040.

[8] J. Sander, M. Ester, H. Kriegel, and X. Xu, "Density-based clustering in spatial databases: The algorithm GDBSCAN and its application", *Data Mining and Knowledge Discovery*, 2, 1998, pp. 169-194.

[9] G. Singh, S. Rajagopalan, and B. Lindsay, "Random sampling techniques for space efficient of large data sets", *Proceedings of ACM SIGMOD Int. Conf. on Management of Data*, 1999.

[10] H. Toivonen, "Sampling large databases for association rules", *Proceedings of International Conference on Very Large Data Bases*, 1996, pp. 134–145.

[11] J. Vitter, "Random sampling with a reservoir", *ACM Transactions on Mathematical Software*, 11, 1985, pp. 37-57.

[12] B. Zhang, "Generalized k-harmonic means--boosting in unsupervised learning", *Technical Report HPL-2000-137*. Hewlett-Packard Labs, 2000.

[13] G. Zipf, *Human Behavior and Principle of Least Effort: An Introduction to Human Ecology*. Addison Wesley, Cambridge, MA, 1949.

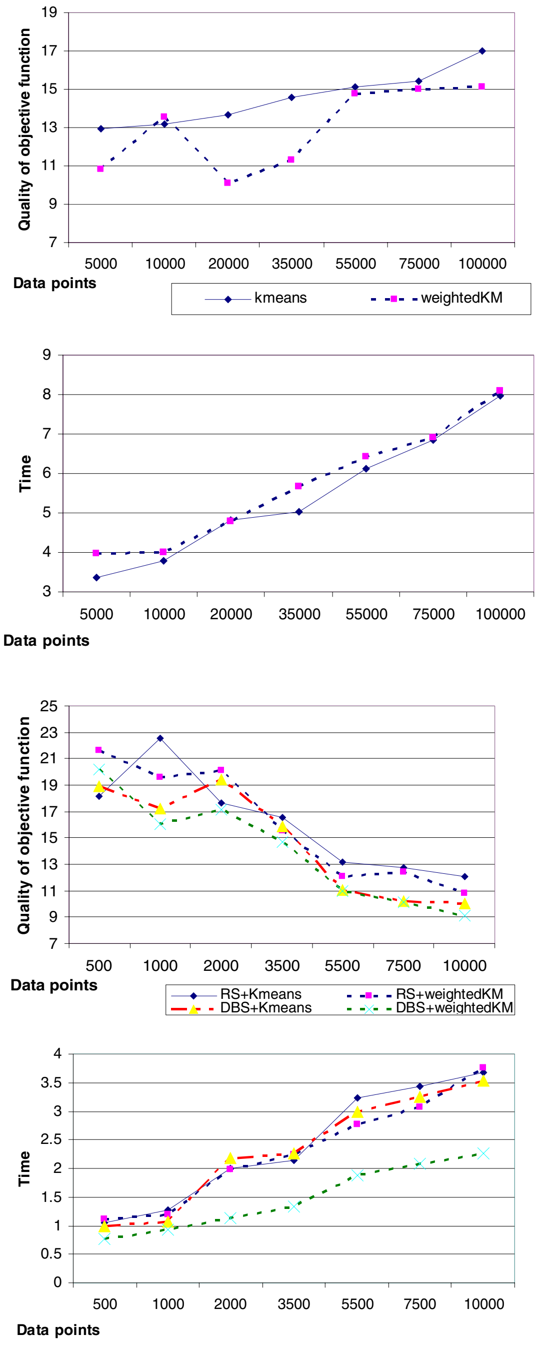


Figure 5. Performance comparison of weighted k-means against k-means and the running time comparison, results on top are experiments running on the whole data set while results at the bottom obtained from running on the sample data