NATIONAL INSTITUTE OF TECHNOLOGY ROURKELA

Seminar and Technical Writing-I

Department of Computer Science & Engineering

A Report on

“A Neighborhood-Based Clustering Algorithm”

Submitted By:
Raj Kapoor Gupta
115CS0016

Submitted To:
Prof. Suchismita Chinara
(Dept. of CSE)

Abstract.

I am very thankful to my college for giving me an opportunity to give a presentation on Neighborhood Based clustering. In this paper,a new clustering algorithm, NBC, i.e., Neighborhood Based Clustering, which discovers clusters based on the neighborhood characteristics of data. The NBC algorithm has the following advantages: (1) NBC is effective in discovering clusters of arbitrary shape and different densities; (2) NBC needs fewer input parameters than the existing clustering algorithms; (3) NBC can cluster both large and high-dimensional databases efficiently.

1 Introduction.

As one of the most important methods for knowledge discovery in databases (KDD), clustering is very useful in many data analysis scenarios, including data mining, document retrieval, image segmentation, and pattern classification [1]. Roughly speaking, the goal of a clustering algorithm is to group the objects of a database into a set of meaningful clusters each of which contains objects as similar as possible according to a certain criterion. Currently, mainly four types of clustering algorithms have been developed, including hierarchical, partitioning, density-based and grid-based algorithms. With the fast development of data collection and data management technologies, the amount of data stored in various databases increases rapidly. Furthermore, more and more new types of data come into existence, such as image, CAD data, geographic data, and molecular biology data. The hugeness of data size and the variety of data types arise new and challenging requirements for clustering algorithms. Generally, a good clustering algorithm should be Effective(e.g. be able to discover clusters of arbitrary shape and different distributions), Efficient(e.g. be able to handle either very large databases or high-dimensional data-bases, and Easy to use(e.g. need no or few input parameters).

The NBC algorithm uses the neighborhood relationship among data objects to build a neighborhood based clustering model to discover clusters. The core concept of NBC is the Neighborhood Density Factor (NDF in abbr.). NDF is a measurement of relative local density, which is quite different the absolute global density used in DBSCAN [2]. In this sense, NBC can still be classified into density based clustering algorithms. However, comparing with DBSCAN(the pioneer and representative of the existing density based clustering algorithms), NBC boasts of the following advantages:

- NBC can automatically discover clusters of arbitrary distribution, it can also recognize clusters of different local-densities and multi-granularities in one dataset, while DBSCAN uses global parameters, it can not distinguish small, close and dense clusters from large and sparse clusters. In this sense, NBC is closer to the Effective criterion than DBSCAN.
- NBC needs only one input parameter(the k value), while DBSCAN requires three input parameters(the k value, the radius of the neighborhood, and the density threshold). That is, NBC needs fewer input parameters than DBSCAN, so NBC is advantageous over DBSCAN in view of the Easy to Use criterion.
- NBC uses cell-based structure and VA file [3] to organize the targeted data, which makes it be efficient and scalable even for very large and high dimensional databases.

With all these advantages, we do not intend to replace the existing algorithms with NBC. Instead, we argue that NBC can be a good complement to the existing clustering methods**.**

2 A Novel Algorithm for Data Clustering

Basic Concepts

The key idea of neighborhood-based clustering is that: for each object p in a cluster, the number of objects whose k-nearest-neighborhood contains p should A Neighborhood-Based Clustering Algorithm 363 not less than the number of objects contained in p’s k-nearest-neighborhood. In what follows, we give the formal definition of neighborhood-based cluster and its related concepts.

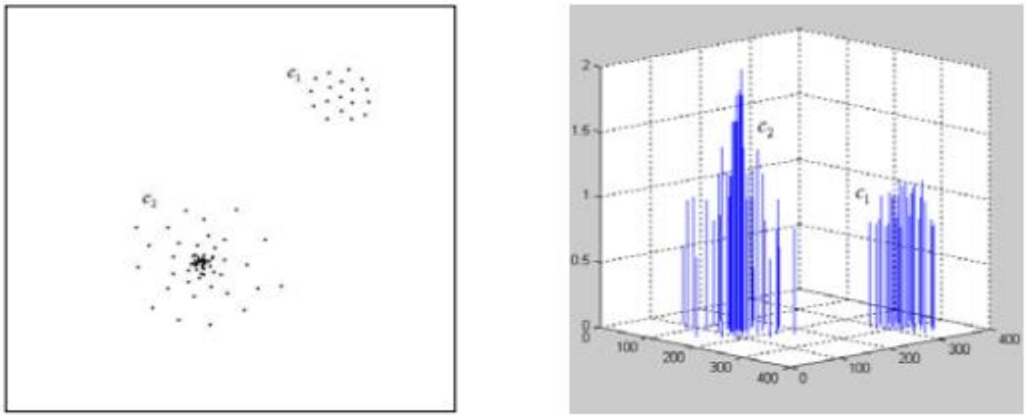
- **Definition 1** (k-Nearest Neighbors Set, or simply kNN). The k-nearest neighbors set of p is the set of k (k > 0) nearest neighbors of p, denoted by kNN(p). In other words, kNN(p) is a set of objects in D such that

(a) |kNN(p)| = k;

(b) p /∈ kNN(p);

(c) Let o and o’ be the k-th and the (k+1)-th nearest neighbors of p respectively, then dist(p, o’) ≥ dist(p, o) holds.

- **Definition 2** (Neighborhood-based Density Factor, or simply NDF). The NDF of point p is evaluated as follows: NDF(p) = |R−kNB(p)|/ |kNB(p)| .



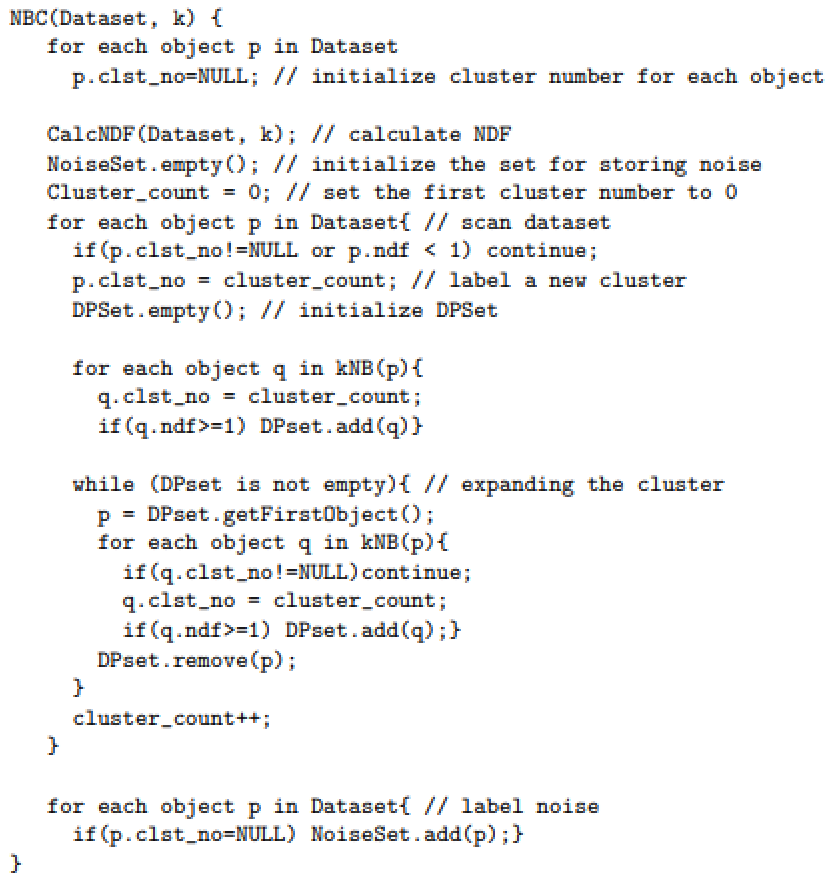
(a) Dataset

(b) NDF values of data points

Fig. 1. An illustration of NDF

- **Definition 3** (Local Dense Point, simply DP). Object p is a local dense point if its NDF(p) is greater than 1, also we call p a dense point w.r.t. kNB(p), denoted by DP w.r.t. kNB(p). The larger NDP(p) is, the denser p’s k-neighborhood is.
- **Definition 4** (Neighborhood-based cluster). Given a dataset D, a cluster C w.r.t. k is a non-empty subset of D such that (a) for two objects p and q in C, p and q are ND-connected w.r.t. k, and (b) if p ∈ C and q is ND-connected from p w.r.t. k, then q ∈ C. The definition above guarantees that a cluster is the maximal set of NDconnected objects w.r.t. k.

The NBC Algorithm:

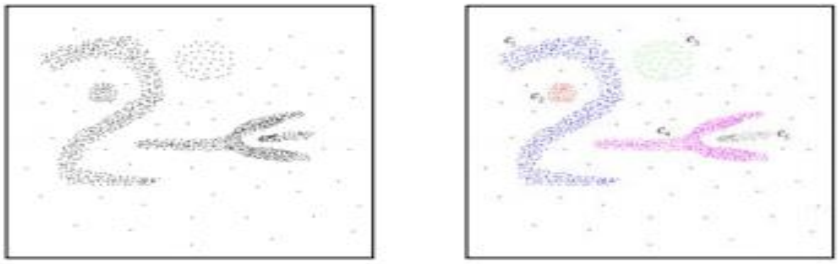


```
NBC(Dataset, k) {
   for each object p in Dataset
     p.clst_no=NULL; // initialize cluster number for each object

   CalcNDF(Dataset, k); // calculate NDF
   NoiseSet.empty(); // initialize the set for storing noise
   Cluster_count = 0; // set the first cluster number to 0
   for each object p in Dataset{ // scan dataset
     if(p.clst_no!=NULL or p.ndf < 1) continue;
     p.clst_no = cluster_count; // label a new cluster
     DPSet.empty(); // initialize DPSet

     for each object q in kNB(p){
       q.clst_no = cluster_count;
       if(q.ndf>=1) DPset.add(q)}

     while (DPset is not empty){ // expanding the cluster
       p = DPset.getFirstObject();
       for each object q in kNB(p){
         if(q.clst_no!=NULL)continue;
         q.clst_no = cluster_count;
         if(q.ndf>=1) DPset.add(q);}
       DPset.remove(p);
     }
     cluster_count++;
   }

   for each object p in Dataset{ // label noise
     if(p.clst_no=NULL) NoiseSet.add(p);}
}
```

Fig. 2. The NBC algorithm in C pseudo-code

Performance Evaluation:

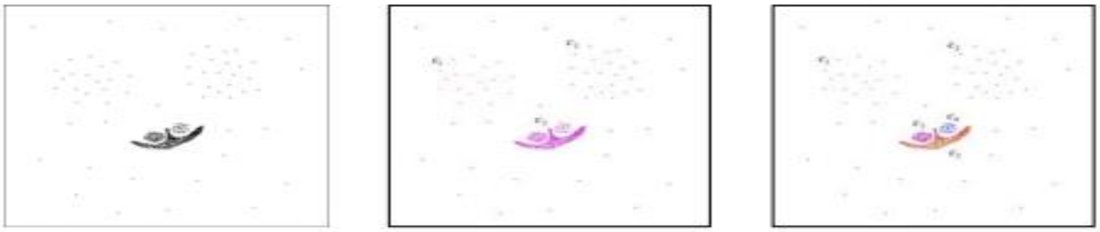
In this section, we evaluate the performance of the NBC algorithm, and compare it with DBSCAN. In the experiments, we take k=10. Considering that k value mainly affect the minimal cluster to find, we do not give the clustering results for different k values. To test NBC's capability of discovering clusters of arbitrary shape, we use a synthetic dataset that is roughly similar to the database 3 in [2], but more complicated. In our dataset, there are five clusters and the noise percentage is 5%. The original dataset and the clustering result of NBC are shown in Fig.3. As is shown, NBC discovered all clusters and recognized the noise points.



(a) Original dataset

(b) Clustering result by NBC

Fig. 3. Discoverying clusters of arbitrary shape

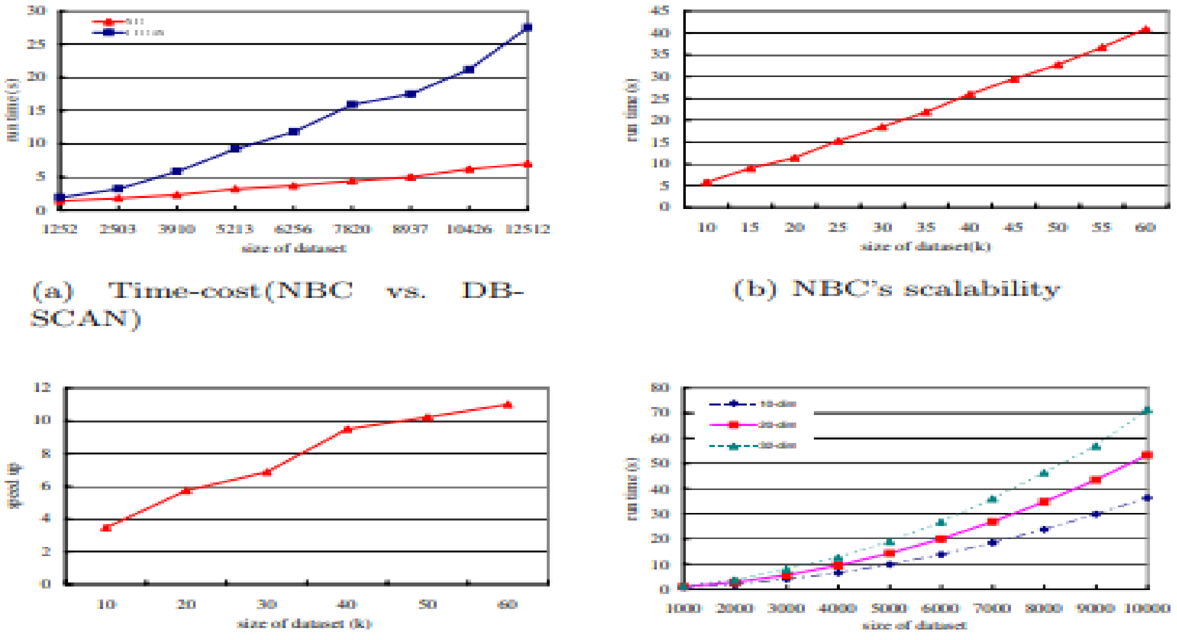


(a) Dataset

(b) DBSCAN's result

(c) NBC's result

Fig. 3. Discoverying clusters of arbitrary shape



(a) Time-cost(NBC vs. DB-SCAN)

(b) NBC's scalability

(c) NBC's Speedup over DB-SCAN

(d) High-dimensional data

Fig. 4. NBC clustering efficiency and scalability:

Conclusion: A new clustering algorithm, NBC, i.e., Neighborhood Based Clustering, which discovers clusters based on the neighborhood relationship among data. It can discover clusters of arbitrary shape and different densities. Experiments show that NBC outperforms DBSCAN in both clustering effectiveness and efficiency. More importantly, NBC needs fewer input parameter from the users than the existing methods.

References

1. J. Han and M. Kamber. Data mining: concepts and techniques. Morgan Kaufmann Publishers, 2000.

2. Ester M., Kriegel H., Sander J., and Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proc. KDD'96, pages 226-231, Portland, Oregon, 1996.

3. R. Weber, H.-J. Schek and S. Blott. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In Proc. of VLDB'98, pages 194-205, New York City, NY, August 1998.

4. C. Merz, P. Murphy, and D. Aha. UCI Repository of Machine Learning Databases. At http://www.ics.uci.edu/ mlearn/MLRepository.html.