Turn Waste into Wealth: On Simultaneous Clustering and Cleaning over Dirty Data

Shaoxu Song Chunping Li Xiaoquan Zhang

KLiss, MoE; TNList; School of Software, Tsinghua University, China
{sxsong, cli, zxq8984}@tsinghua.edu.cn

ABSTRACT

Dirty data commonly exist. Simply discarding a large number of inaccurate points (as noises) could greatly affect clustering results. We argue that dirty data can be repaired and utilized as strong supports in clustering. To this end, we study a novel problem of clustering and repairing over dirty data at the same time. Referring to the minimum change principle in data repairing, the objective is to find a minimum modification of inaccurate points such that the large amount of dirty data can enhance the clustering. We show that the problem can be formulated as an integer linear programming (ILP) problem. Efficient approximation is then devised by a linear programming (LP) relaxation. In particular, we illustrate that an optimal solution of the LP problem can be directly obtained *without calling a solver*. A quadratic time approximation algorithm is developed based on the aforesaid LP solution. We further advance the algorithm to linear time cost, where a trade-off between effectiveness and efficiency is enabled. Empirical results demonstrate that *both the clustering and cleaning accuracies* can be improved by our approach of repairing and utilizing the dirty data in clustering.

Categories and Subject Descriptors

I.5.3 [**Clustering**]: Algorithms

Keywords

Data repairing; data cleaning

1. INTRODUCTION

Density-based clustering can successfully identify noises (see a survey in [16]). However, rather than a small proportion of noise points, real data are often dirty with a large number of inaccurate points [10]. For instance, a (very) large portion of GPS data are inaccurate, especially in the indoor environment with weak signals. According to our experiments (in Section 6), 139 out of 818 (about 17%) GPS readings are inaccurate. Similar examples of inaccurate data include RFID sensor readings [4] or multimedia data [11].

The large amount of noise points are simply discarded, if we directly apply the existing density-based clustering approaches, e.g., the well-known DBSCAN [7]. With too much information loss, clustering results could be dramatically affected (see examples below).

Instead of discarding dirty data, we argue that the large amount of dirty data can be repaired and utilized as strong supports in clustering. To the best of our knowledge, this is the first study on performing data cleaning and clustering at the same time.

A natural idea is to first clean the dirty data before clustering. Existing constraint-based cleaning techniques can be applied, e.g., repairing with *functional dependencies* (FD) [3] (see [8] for a survey). According to our empirical study (in Figure 11), the clustering accuracy can be improved by applying first the FD-based repairing, when the FD constraints are available in the considered dataset. However, for some other datasets with simple schema, such as GPS data, no such (FD) integrity constraints could be declared (and thus the constraint-based repairing is not applicable).

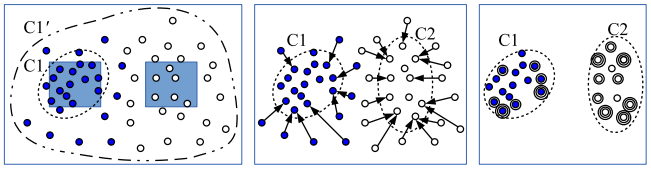
Besides repairing mentored by integrity constraints, we propose to repair the dirty data under the guidance of *density* information, inspired by the successful identification of noisy data in density-based clustering. The idea is to simultaneously repair dirty data w.r.t. the density of data during the clustering process, rather than separately repairing (ahead) w.r.t. integrity constraints. By interchangeably taking the advantages of density-based clustering and repairing, *both the clustering and repairing tasks benefit* (with accuracy improvement as shown in the experiments in Section 6).

Following the minimum change principle in data repairing [24], i.e., the change made in repairing is expected to be as small as possible, the objective of simultaneous repairing and clustering is to find a minimum repair of data such that all the data can be clustered (utilized). The rationale of minimum change is that systems or humans always try to minimize mistakes in practice. That is, dirty values, such as inaccurate GPS readings or typos in text, are often not very far from true values. Referring to this objective, we formalize the problem as an optimization problem, namely *Density-based Optimal Repairing and Clustering* (DORC).

Example 1. We illustrate a motivation example of GPS data in Figure 1. Consider GPS readings in two nearby buildings, denoted by blue and white points in Figure 1(a), respectively. C1 denotes more precise points (high-density, in a building with good GPS signal), whereas C2, for a

KDD'15, August 10-13, 2015, Sydney, NSW, Australia.

DOI: http://dx.doi.org/10.1145/2783258.2783317.



(a) clusters in dirty data (b) repairing and clustering (c) repaired data

Figure 1: Clustering with repairing over dirty data

building with weaker GPS signal, contains inaccurate points (low-density, that drop far away from the building and need concentrating repairing). The density-based clustering such as DBSCAN either returns a cluster C1 (with high density requirements) or C1′ (with low density parameters). White points corresponding to the second physical building cannot form a separate cluster (either directly ignored as noisy data in C1 or merged with blue ones in C1′). The massive points identified as noises by DBSCAN in Figure 1(a) are strong evidence to concentrate C2. (See more example results in Figure 6 in Section 6.1)

In this study, we propose to repair the inaccurate data points during the clustering process. For example, as shown in Figure 1(b), an arrow ($a \to b$) denotes that a point is repaired from location a to location b. Since points are concentrated (with higher density) after repairing, two clusters are successfully formed in Figure 1(c).

The problem of clustering with repairing, however, is nontrivial. Simply repairing noise points to the closest clusters is not sufficient, e.g., repairing all the noise points to C1 in Figure 1 does not help in identifying the second cluster C2. Indeed, it should be considered that dirty points may possibly form clusters with repairing (i.e., C2). □

It is notable that our proposed DORC techniques are complementary to the existing constraint-based repairing (when constraints are available). As illustrated in experiments (Figure 12 in Section 6.3), by combining our DORC approach with the existing FD-based repairing, both the clustering and repairing accuracies are further improved. Compared to existing constraint-based repairing, the major advantages of DORC are in two aspects: (1) it does not require any external knowledge of integrity constraints or rules; (2) instead, it explores the density information embedded inside the data, which is not considered in the preliminary constraint-based methods. In this sense, our formulation favors errors with significant distortion on density, while minor errors without altering density might not be handled.

Our major contributions in this paper are summarized as:

(1) We formalize the DORC problem of simultaneous clustering and repairing. In particular, *no additional parameters* are introduced for DORC besides the density and distance requirements η and ε for clustering.

(2) We formulate DORC as an ILP problem (in Section 3), investigate its LP relaxation and, most importantly, show that an optimal solution to the LP problem can be *directly obtained without calling a solver*.

(3) We devise a *quadratic time* approximation algorithm QDORC upon the LP solution (in Section 4).

(4) We advance the algorithm to *linear time* complexity (in Section 5). The more efficient LDORC supports a *trade-off* between effectiveness and efficiency, via a group distance threshold τ.

Table 1: Notations

Symbol	Description
n	$\lvert\mathcal{P}\rvert$ the number of data points $p \in \mathcal{P}$
m	$\lvert\mathcal{L}\rvert$ the number of leader points $p \in \mathcal{L}$
δ	distance of data points, $\delta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_0^+$
ε	distance threshold, Eps
η	density threshold, MinPts
h_{ij}	ε-neighborhood of two points p_i to p_j
λ	repair of data points, $\lambda : \mathcal{P} \to \mathcal{P}$
$\Delta(\lambda)$	cost of a repair λ
$C(p_j)$	set of ε-neighbors of point p_j
c_j	count of neighbors in location p_j after repairing
τ	group distance threshold

(5) We report an extensive experimental study on both real and synthetic datasets (in Section 6). The results demonstrate that *both the clustering and repairing accuracies* are improved by our proposed DORC approaches.

Table 1 lists the frequently used notations in this paper. Proofs of all lemmas and propositions can be found in [1].

2. PROBLEM STATEMENT

Clustering. Consider a set of data points $\mathcal{P}$. Let $\delta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_0^+$ be a distance function, satisfying nonnegativity $\delta(p_i, p_j) \geq 0$, identity of indiscernibles $\delta(p_i, p_j) = 0$ iff $p_i = p_j$, symmetry $\delta(p_i, p_j) = \delta(p_j, p_i)$, where $p_i, p_j \in \mathcal{P}$.

Two points $p_i, p_j \in \mathcal{P}$ are said to be in ε-neighborhood, if $\delta(p_i, p_j) \leq \varepsilon$. We denote $C(p_i) = \{p_j \in \mathcal{P} \mid \delta(p_i, p_j) \leq \varepsilon\}$ the set of ε-neighbors of p_i, where $p_i \in C(p_i)$ as well.

Definition 1 (Core points, Border points, Noise points)**.** *Given a distance threshold ε (Eps) and a density threshold η (called MinPts), a point p_i with $\lvert C(p_i)\rvert \geq \eta$ is considered as a* core point*. A* border point *has $\lvert C(p_i)\rvert < \eta$ but is in ε-neighborhood of some core point. All the other points, which are neither core points nor border points (in ε-neighborhood of some core point), are* noise points*.*

Repairing. A *repair* over a set of points is a mapping $\lambda : \mathcal{P} \to \mathcal{P}$. We denote $\lambda(p_i)$ the location of point p_i after repairing. The ε-neighbors of $\lambda(p_i)$ after repairing is $C_\lambda(p_i) = \{p_j \in \mathcal{P} \mid \delta(\lambda(p_i), \lambda(p_j)) \leq \varepsilon\}$

Following the minimum change principle in data cleaning that we prefer a repair close to the input [24], the repairing cost $\Delta(\lambda)$ is defined as

$$\Delta(\lambda) = \sum_{i=1}^{n} w(p_i, \lambda(p_i)), \tag{1}$$

where $w(p_i, \lambda(p_i))$ is the cost of repairing a point p_i to the new location $\lambda(p_i)$. For instance, a count cost [15], with $w(p_i, \lambda(p_i)) = 1$ iff $p_i \neq \lambda(p_i)$, considers the number of data points that are modified as the repairing cost. Alternatively, the distance of point locations before and after repairing could also be considered [3], with $w(p_i, \lambda(p_i)) = \delta(p_i, \lambda(p_i))$.

DORC Problem. As mentioned in the introduction, by simply relaxing parameters in DBSCAN, the diffusion of nearby

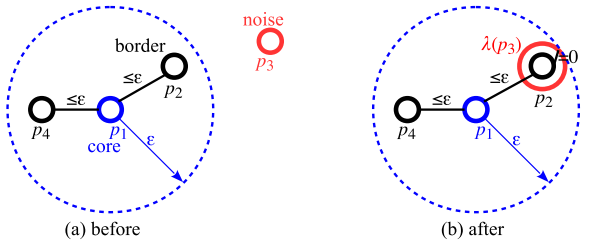


Figure 2: Example of repairing

clusters may force them to combined, e.g., in Figure 1(a), C2 (white points) is either ignored as noises by C1 or merged as C1' (by relaxing the density parameter). We propose to utilize the dirty (noise) points for clustering by repairing. That is, the noise points are repaired and thus clustered as either core points or border points. In other words, for each repaired $\lambda(p_i)$, either itself or one of its ε-neighbors has a ε-neighborhood size greater than MinPts η. Since a cluster is uniquely determined by its core points [7], the repairing process with identification of core points (in the repair results) outputs the clustering results as well.

Problem 1. *Given a set of data points $\mathcal{P}$, a distance threshold ε and a density threshold η, the* Density-based Optimal Repairing and Clustering *(DORC) problem is to find a repair λ (i.e., a mapping $\lambda : \mathcal{P} \to \mathcal{P}$) such that*

(1) the repairing cost $\Delta(\lambda)$ is minimized, and

(2) for each repaired $\lambda(p_i)$, either $|C_\lambda(p_i)| \geq \eta$ (core points), or $|C_\lambda(p_j)| \geq \eta$ for some p_j with $\delta(\lambda(p_i), \lambda(p_j)) \leq \varepsilon$.

It is worth noting that multiple points may be repaired to the same "physical" location, having $\lambda(p_i) = \lambda(p_j)$.

Example 2. Consider a clustering density requirement $\eta = 3$. As shown in Figure 2(a), point p_1, whose $|C(p_1)| = |\{p_1, p_2, p_4\}| = 3$, is a core point. Points p_2 and p_4 in ε-neighborhood of p_1 are border points. Point p_3, not in ε-neighborhood of any point, is considered as a noise point.

Figure 2(b) illustrates a possible repair λ, where p_3 is moved to the location of p_2, i.e., $\lambda(p_3) = p_2$. Point p_2 with $\lambda(p_2) = p_2$ remains unchanged (and similarly for p_1, p_4). There are two points in the location of p_2 after repairing (denoted by red and black concentric circles). We have $C_\lambda(p_2) = C_\lambda(p_3) = \{p_1, p_2, p_3\}$. That is, points p_2 and p_3 upgrade to core points by repairing. □

3. ILP FORMULATION

In this section, we illustrate how to formulate the DORC problem as an ILP problem, and thus existing ILP solvers can be directly applied (a built-in advantage).

Consider variable $x_{ij}, 0 \leq x_{ij} \leq 1$. Let $x_{ij} = 1$ denote that point p_i is repaired to location p_j after repairing, i.e., $\lambda(p_i) = p_j$; otherwise, $x_{ij} = 0$. Obviously, a point can only be repaired to one location, having

$$\sum_{j=1}^{n} x_{ij} = 1. \tag{2}$$

The weight w_{ij} for x_{ij} is defined as the corresponding cost of repairing p_i to p_j, $w_{ij} = w(p_i, p_j)$.

After repairing, there may exist multiple points p_i being repaired to the location of a point p_j. The new ε-neighborhood count of location p_j is

$$c_j = |\{p_i \in \mathcal{P} \mid \delta(\lambda(p_i), p_j) \leq \varepsilon\}| = \sum_{i=1}^{n} x_{ij} + \sum_{k=1}^{n}\sum_{i=1}^{n} h_{jk}x_{ik}, \tag{3}$$

where $h_{jk} = 1$ denotes that locations p_j and p_k are in ε-neighborhood; otherwise, $h_{jk} = 0$. That is, c_j counts the total number of points located in p_j and all of its ε-neighbors p_k, after repairing.

Let $y_j = 1$ denote that location p_j has ε-neighbor count no less than η, i.e., core location; otherwise, $y_j = 0$. It follows

$$\frac{c_j}{\eta} \geq y_j \geq \frac{c_j - \eta + 1}{n}, \tag{4}$$

where $n = |\mathcal{P}|$ is the total number of points. It specifies that $y_j = 1$ iff $c_j \geq \eta$; and $y_j = 0$ iff $c_j < \eta$.

The repairing should ensure eliminating all noise points. In other words, a point is either a core point or a border point (which is a neighbor of a core point). More precisely, for any location p_j with at least one point retained after repairing, i.e., $x_{kj} = 1$ for some k, it is required that either this point or one of its neighbors belongs to core points (with ε-neighborhood size no less than η). We have

$$y_j + \sum_{i=1}^{n} y_i h_{ij} \geq \frac{1}{n}\sum_{k=1}^{n} x_{kj}. \tag{5}$$

In other words, for any j with $x_{kj} = 1$ for some k, it requires either $y_j = 1$ or some other $y_i = 1$ such that p_i is in ε-neighborhood with p_j ($h_{ij} = 1$).

Given the constraints in formulas (2), (4) and (5), the DORC problem is formulated as the following ILP problem.

$$\begin{aligned}
\text{minimize} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_{ij} \\
\text{subject to} \quad & \sum_{j=1}^{n} x_{ij} = 1, && 1 \leq i \leq n \\
& c_j - \eta y_j \geq 0, && 1 \leq j \leq n \\
& y_j n - c_j \geq 1 - \eta, && 1 \leq j \leq n \\
& y_j + \sum_{i=1}^{n} y_i h_{ij} - \frac{1}{n}\sum_{k=1}^{n} x_{kj} \geq 0, && 1 \leq j \leq n \\
& x_{ij}, y_j \in \{0, 1\} \qquad 1 \leq i \leq n, && 1 \leq j \leq n
\end{aligned} \tag{6}$$

Existing ILP solvers can be directly applied to compute the optimal solutions. It returns not only a repair x_{ij} but also a set of core points $\lambda(p_i) = p_j$ with $y_j = 1$ after repairing.

Proposition 1. *The optimal solution $\mathbf{x}^{\text{ILP}}, \mathbf{y}^{\text{ILP}}$ of ILP forms an optimal repair λ^{ILP} with the minimum repairing cost*

$$\Delta(\lambda^{\text{ILP}}) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_{ij}^{\text{ILP}},$$

where $\lambda^{\text{ILP}}(p_i) = p_j$ iff $x_{ij}^{\text{ILP}} = 1, 1 \leq i \leq n, 1 \leq j \leq n$.

Example 3 (Example 2 continued)**.** Consider again the example of 4 points in Figure 2 with clustering density requirement $\eta = 3$. We show how the repair λ, with $\lambda(p_3) = p_2$

in Figure 2(b), corresponds to the feasible solution $\mathbf{x}$ to the ILP, where $x_{11} = x_{22} = x_{44} = 1$ and $x_{32} = 1$.

For the location of p_2, we have $c_2 = 2 + 1 = 3$ by formula (3). It follows $y_2 = 1$ according to formula (4). In other words, all the points in the location of p_2 after repairing are core points, i.e., p_2 and p_3 as indicated in Example 2.

For the location of p_3, since there is no point retained after repairing, having $\sum_{k=1}^{4} x_{k3} = 0$, formula (5) is satisfied. Therefore, the solution corresponding to the repair λ satisfies all the constraints in formula (6) and is a feasible solution of ILP. □

We can show that the DORC problem is always solvable.

Proposition 2. *For $\eta < n$, a feasible solution to the* ILP *problem always exists.*

Proof. By simply repairing all the points to a single location say p_1, we have $x_{i1} = 1$ and $y_1 = 1$, i.e., all the points become core points locating in p_1 after repairing. □

4. QUADRATIC TIME APPROXIMATION

Another advantage of modeling DORC as ILP is the possible LP relaxation for efficiently computing near optimal solutions. In this section, we first indicate that an optimal LP solution can be directly derived without calling a solver. An approximation algorithm is then built upon the LP solution.

4.1 LP Solution without Calling a Solver

The ILP problem can be relaxed as a LP problem by changing the integer constraints in formula (6) to $0 \leq x_{ij} \leq 1, 0 \leq y_j \leq 1$. We show below that an optimal solution to the LP problem can be directly obtained without calling a solver.

Lemma 3. *There always exists an optimal solution $\mathbf{x}^{\text{LP}}, \mathbf{y}^{\text{LP}}$ for the* LP *problem, where*

$$x_{ii}^{\text{LP}} = 1, \quad i = 1, \ldots, n \tag{7}$$

$$x_{ij}^{\text{LP}} = 0, \quad i \neq j, i = 1, \ldots, n, j = 1, \ldots, n$$

$$y_j^{\text{LP}} = \begin{cases} 1 & \text{if } \sum_{k=1}^{n} h_{jk} \geq \eta \\ \frac{\sum_{k=1}^{n} h_{jk}}{\eta} & \text{if } \sum_{k=1}^{n} h_{jk} < \eta \end{cases} \quad j = 1, \ldots, n$$

4.2 Clustering&Repairing with LP Solution

As mentioned, a cluster is uniquely determined by its core point [7]. The clustering and repairing process is thus to round the LP solution, i.e., to determine $x_{ij} \in \{0, 1\}, y_j \in \{0, 1\}$ according to the aforesaid LP solution $(\mathbf{x}^{\text{LP}}, \mathbf{y}^{\text{LP}})$.

Before introducing the algorithm, let us first rephrase core, border and noise points in Definition 1 in terms of neighborhood $\mathbf{h}$.

Definition 2 (Core points, Border points, Noise points)**.** *For a set of data points $\mathcal{P}$ and the corresponding neighborhood $\mathbf{h}$, core points $\mathcal{C}(\mathbf{h})$, border points $\mathcal{B}(\mathbf{h})$, and noise points $\mathcal{N}(\mathbf{h})$ w.r.t. $\mathbf{h}$ are*

$$\mathcal{C}(\mathbf{h}) = \{p_i \in \mathcal{P} \mid (\textstyle\sum_{j=1}^{n} h_{ij}) \geq \eta\},$$

$$\mathcal{B}(\mathbf{h}) = \{p_i \in \mathcal{P} \mid (\textstyle\sum_{j=1}^{n} h_{ij}) < \eta, h_{ik} = 1, p_k \in \mathcal{C}(\mathbf{h})\},$$

$$\mathcal{N}(\mathbf{h}) = \mathcal{P} \setminus (\mathcal{C}(\mathbf{h}) \cup \mathcal{B}(\mathbf{h})).$$

Algorithm 1 (QDORC) presents an approximation to DORC. We consider all the noise points $\mathcal{N}(\mathbf{h})$ w.r.t. $\mathbf{h}$, denoted by $\mathcal{N}$ in Line 1. Note that y_j^{LP} in the LP solution can be interpreted as the probability of a point p_j being core point in clustering. In each step, we consider a point $p_j \in \mathcal{P}$ with the largest y_j^{LP} in Line 4. In order to become a core point, point p_j needs at least $(1 - y_j)\eta$ additional ε-neighbors. We heuristically consider the noise point p_i with the minimum w_{ij} (in Line 7) and conduct the repairing of p_i to p_j, i.e., assign $x_{ij} = 1$. If there are no sufficient noise points retained for repairing, i.e., the remaining noise points cannot form at least η neighbors of the point p_j, we repair all the remaining noise points to their closest non-noise points in Line 13.

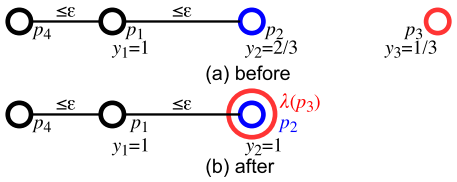


Figure 3: Example of repairing with LP solution

Algorithm 1: QDORC$(\mathcal{P}, \mathbf{h}, \varepsilon, \eta)$

Data: A set of data points $\mathcal{P}$ with neighborhood $\mathbf{h}$, distance threshold ε and density threshold η
Result: A set of core locations with $y_j = 1$ in $\mathbf{y}$ and the corresponding repairing x_{ij} in $\mathbf{x}$

1 $\mathcal{N} := \mathcal{N}(\mathbf{h})$ w.r.t. $\mathbf{h}$;
2 $\mathbf{x}, \mathbf{y} := \mathbf{x}^{\text{LP}}, \mathbf{y}^{\text{LP}}$ the LP solution w.r.t. $\mathbf{h}$ in Lemma 3;
3 **while** $\mathcal{N} \neq \emptyset$ **do** // noise points exist
4 let $p_j \in \mathcal{P}$ with the maximum y_j and $y_j < 1$;
5 **if** $|\mathcal{N}| \geq (1 - y_j)\eta$ **then**
6 **repeat**
7 let $p_i \in \mathcal{N}$ with the minimum w_{ij};
8 $x_{ii} := 0, x_{ij} := 1$ and $\mathcal{N} := \mathcal{N} \setminus \{p_i\}$;
9 **until** $(1 - y_j)\eta$ *times*;
10 $y_j := 1$ and $\mathcal{N} := \mathcal{N} \setminus (\{p_j\} \cup \{p_k \mid h_{jk} = 1\})$;
11 **else** // no sufficient noises retain
12 **for** *each* $p_i \in \mathcal{N}$ **do**
13 let $p_k \in \mathcal{P} \setminus \mathcal{N}$ with the minimum w_{ik};
14 $x_{ii} := 0, x_{ik} := 1$ and $\mathcal{N} := \mathcal{N} \setminus \{p_i\}$;
15 **return** $\mathbf{y}, \mathbf{x}$

The returned result $\mathbf{x}$ corresponds to a repair λ^{QDORC} such that $\lambda^{\text{QDORC}}(p_i) = p_j$ for $x_{ij} = 1, 1 \leq i \leq n, 1 \leq j \leq n$.

Example 4. Consider an example of 4 points in Figure 3(a) with clustering density requirement $\eta = 3$. Each edge, e.g., (p_1, p_2), denotes that two points are in ε-neighborhood. Point p_3 is identified as a noise point, referring to Definition 2 w.r.t. neighborhoods, having $\mathcal{N} = \{p_3\}$. According to the formulas in Lemma 3, we derive an LP solution, with $y_1 = 1, y_2 = 2/3, y_3 = 1/3, y_4 = 2/3$.

Line 3 in Algorithm 1 selects a point p_2 with the maximum $y_2 = 2/3 < 1$. To make p_2 a core point, i.e., $y_2 = 1$, we need at least one additional noise point, to repair to the location of p_2. Line 7 selects $p_3 \in \mathcal{N}$, and repair it to the location of p_2 by setting $x_{32} = 1$.

Since no noise point retains in $\mathcal{N}$, the algorithm terminates and returns a solution with $x_{11} = x_{22} = x_{44} = 1, x_{32} = 1, y_1 = y_2 = 1$ (all the others are equal to 0). It corresponds to a repair λ with $\lambda(p_3) = p_2$ (and 3 other unchanged points)

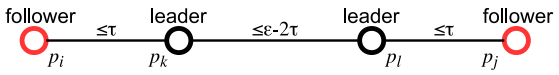


Figure 4: Example of leaders and followers

and a set of core points $\{\lambda(p_1), \lambda(p_3), \lambda(p_3)\}$ (located in the core locations of p_1, p_2). □

Proposition 4. *Algorithm 1 (QDORC) runs in* $O(n^2)$ *time, returns a feasible solution to the* ILP *problem, where* $n = |\mathcal{P}|$.

5. LINEAR TIME APPROXIMATION

It is worth noting that capturing the ε-neighborhood $\mathbf{h}$ for QDORC is costly (in $O(n^2)$ time). Following the same line of performing efficient density-based clustering [23], we present an algorithm for efficiently estimating the neighborhood by partitioning data points into groups, and perform DORC over the estimate neighborhood (in linear time).

5.1 Estimating Neighborhood via Grouping

Let τ be a distance threshold in grouping, $0 \leq \tau \leq \frac{\varepsilon}{2}$. Each group consists of a leader data point p and a set of follower data points, denoted by $\mathsf{follower}(p)$. Each follower should have distance to its leader no greater than the group distance threshold τ, i.e., $\delta(p, p_i) \leq \tau, \forall p_i \in \mathsf{follower}(p)$.

Lines 2-8 in Algorithm 2 present the grouping procedure of generating leaders and followers. Let $\mathcal{L}$ denote the set of leaders. Each point $p \in \mathcal{P}$ is either assigned as a follower of some existing leader $p_l \in \mathcal{L}$ (in Line 8 if $\delta(p_l, p) \leq \tau$), or created as a new leader (in Line 5).

We introduce approximate ε-neighborhoods, $\mathbf{h}^L$ for estimating $\mathbf{h}$. As presented in Line 11 in Algorithm 2, for two leaders $p_k, p_l \in \mathcal{L}$, we assign $h^L_{kl} := 1$ if $\delta(p_k, p_l) \leq \varepsilon - 2\tau$ rather than $\delta(p_k, p_l) \leq \varepsilon$ for $h_{kl} := 1$. For any $p_i \in \mathsf{follower}(p_k), p_j \in \mathsf{follower}(p_l)$, as illustrated in Figure 4, we use the neighborhood of leaders to approximate that of followers, by assigning $h^L_{ij} = h^L_{kl}$. Thereby, to count the number of neighbors w.r.t. h^L_{ij} for any follower $p_j \in \mathsf{follower}(p_l)$, it is equivalent to "count" the corresponding leaders.

Lemma 5. *The neighbor count w.r.t.* h^L_{ij} *of a point* p_j *has*

$$\sum_{i=1}^{n} h^L_{ij} = \sum_{k=1}^{m} h^L_{kl} |\mathsf{follower}(p_k)|,$$

where $p_i \in \mathsf{follower}(p_k), p_j \in \mathsf{follower}(p_l)$, *and* $m = |\mathcal{L}|$ *is the number of leaders.*

When calling QDORC$(\mathcal{P}, \mathbf{h}^L, \varepsilon, \eta)$ in Line 12 in Algorithm 2 LDORC, the noise points $\mathcal{N} = \mathcal{N}(\mathbf{h}^L)$ in Line 1 and the LP solution $\mathbf{x}, \mathbf{y}$ in Line 2 in Algorithm 1 QDORC can be computed by counting the aforesaid leaders (in $O(mn)$ time).

For heuristically choosing repair candidates w.r.t. weight w in Lines 7 and 13 of Algorithm 1 QDORC, we employ again the leaders. Similar to approximating the neighborhood of followers by that of leaders, we replace Line 7 of Algorithm 1 by choosing a $p_i \in \mathsf{follower}(p_k)$ such that the weight w_{kl} of leaders is minimized, where $p_j \in \mathsf{follower}(p_l), k \neq l$.

Example 5. Consider the example of 4 points in Figure 5(a). Suppose that the points are processed in an order of p_1, p_2, p_3, p_4 during grouping. Two groups are generated with leaders p_1 and p_3. The followers are $\mathsf{follower}(p_1) = \{p_1, p_2\}$ and $\mathsf{follower}(p_3) = \{p_3, p_4\}$. Suppose that $\delta(p_1, p_3) > \varepsilon - 2\tau$. The estimate neighborhoods are $h^L_{12} = h^L_{21} = h^L_{34} = h^L_{43} = 1$ (all the others are equal to 0). Algorithm 2 then calls QDORC$(\mathcal{P}, \mathbf{h}^L, \varepsilon, \eta)$ in Line 12 to process the repairing.

According to Lemma 5, the number of neighbors w.r.t. $\mathbf{h}^L$ of a follower, say p_2, can be calculated by counting the sizes (number of followers) of all leaders with distance to p_1 (leader of p_2) in $\varepsilon - 2\tau$, i.e., $\sum_{i=1}^{n} h^L_{i2} = |\mathsf{follower}(p_1)| = 2$. The same neighbor count is calculated for the other points.

Given a clustering density threshold $\eta = 4$ and the aforesaid neighbor counts, all the 4 points are identified as noise points w.r.t. $\mathbf{h}^L$, i.e., $\mathcal{N}(\mathbf{h}^L) = \{p_1, p_2, p_3, p_4\}$ referring to Definition 2. The LP solution has $y_1 = y_2 = y_3 = y_4 = 2/4$.

Suppose that p_2 is first considered in Line 4 (in Algorithm 1) when calling QDORC$(\mathcal{P}, \mathbf{h}^L, \varepsilon, \eta)$. Referring to $y_2 = 2/4$, it has to repair at least two noise points in order to make itself a core point. As introduced in the paragraph before Example 5, we consider a point (say p_4) in $\mathsf{follower}(p_3)$ to repair, since p_3 (is the only leader that) has the minimum weight w_{31} to the leader p_1 of p_2. As illustrated in Figure 5(b), p_4 is repaired to the location of p_2, having $x_{42} = 1$ and $\lambda(p_4) = p_2$. The algorithm carries on by repairing p_3 to the location of p_2, $\lambda(p_3) = p_2$, in order to make $y_2 = 1$. Since p_2 upgrades to a core point, its neighbor p_1 is removed from $\mathcal{N}$ as well. We have $\mathcal{N} = \emptyset$ and the algorithm terminates. □

Algorithm 2: LDORC$(\mathcal{P}, \tau, \varepsilon, \eta)$

Data: Data point set $\mathcal{P}$, group distance threshold τ, cluster distance threshold ε, density threshold η
Result: A set of core locations with $y_j = 1$ in $\mathbf{y}$ and the corresponding repairing x_{ij} in $\mathbf{x}$

1 $\mathcal{L} := \emptyset$;
2 **for** *each point* $p \in \mathcal{P}$ **do**
3 find the first leader $p_l \in \mathcal{L}$ s.t. $\delta(p_l, p) \leq \tau$;
4 **if** p_l *does not exist* **then**
5 $\mathcal{L} = \mathcal{L} \bigcup \{p\}$; // create a new leader
6 $\mathsf{follower}(p) := \{p\}$;
7 **else**
8 $\mathsf{follower}(p_l) := \mathsf{follower}(p_l) \bigcup \{p\}$;
9 **for** *each leader pair* $p_k, p_l \in \mathcal{L}$ **do**
10 **if** $\delta(p_k, p_l) \leq \varepsilon - 2\tau$ **then**
11 $h^L_{kl} := 1$; // for neighborhood estimation
12 **return** QDORC$(\mathcal{P}, \mathbf{h}^L, \varepsilon, \eta)$

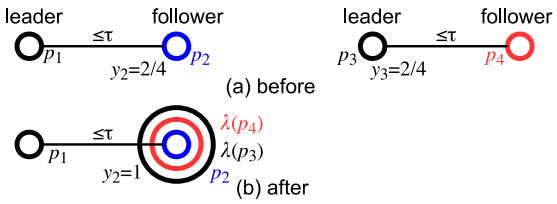


Figure 5: Example of repairing with grouping

Proposition 6. *Algorithm 2 (LDORC) runs in* $O(mn)$ *time, where* $n = |\mathcal{P}|$ *and* $m = |\mathcal{L}|$.

5.2 Performance Analysis

$\mathbf{h}^L$ contains false negatives, i.e., when a $h^L_{ij} = 0$ indicates that the neighborhood does not exist between p_i and p_j

(the result is negative), but it is in fact present ($h_{ij} = 1$). Nevertheless, $\mathbf{h}^L$ will never lead to false positives, where $h^L_{ij} = 1$ but $h_{ij} = 0$.

Lemma 7. *For the neighborhood estimation for* $\mathbf{h}$*, it always has* $h^L_{ij} \leq h_{ij}$.

In other words, the neighborhood $\mathbf{h}^L$ is a subset of $\mathbf{h}$. The returned result w.r.t. $\mathbf{h}^L$ must also be a feasible solution of ILP w.r.t. $\mathbf{h}$. Correctness of Algorithm 2 is guaranteed.

Indeed, $\mathcal{N}(\mathbf{h}^L)$ considered for repairing in LDORC is a superset of $\mathcal{N}(\mathbf{h})$ for the original QDORC.

Lemma 8. *For the noise points w.r.t. the neighborhood, we have* $\mathcal{N}(\mathbf{h}^L) \supseteq \mathcal{N}(\mathbf{h})$.

We show that the number of leaders m is bounded by a constant w.r.t. τ, given a finite domain of data instances (i.e., with a bounded maximum distance of two points).

Proposition 9. *The number of leaders* $m = |\mathcal{L}|$ *is bounded by* $m < (\frac{\delta_{\max}}{\tau} + 1)^2$*, where* $\delta_{\max}$ *is the maximum distance between two points, and* τ *is the group distance threshold.*

Combining Propositions 6 and 9, Algorithm 2 (LDORC) runs in $O((\frac{\delta_{\max}}{\tau} + 1)^2 n)$ time, i.e., a linear time algorithm in the number of data points n. When given a finite space of data instances, the maximum distance of two points $\delta_{\max}$ is a constant. However, for an infinite space, $\delta_{\max}$ could be arbitrarily large.

Trade-off between Effectiveness and Efficiency

Indeed, the parameter τ of group distance threshold provides a trade-off between efficiency and effectiveness.

First, the larger (closer to $\delta_{\max}$) the group distance threshold τ, the smaller the bound of m (number of leaders) is, i.e., lower algorithm time cost. However, for an extremely large τ, e.g., the largest $\tau = \frac{\varepsilon}{2}$, no neighborhood between leaders exist given $\varepsilon - 2\tau = 0$. That is, only the neighborhoods between the points in the same group can be identified in $\mathbf{h}^L$. Such a weak estimation leads to inaccurate computation of $\mathbf{y}$ in the LP solution, and thus the corresponding repairing might not be reliable (i.e., lower repairing and clustering accuracy as illustrated in Figure 10 of experiments).

On the other hand, for a small τ, e.g., $\tau = 0$, it leads to single size groups, where each point corresponds to the leader of a group but without any other follower. In this sense, LDORC is exactly QDORC without grouping. Since the LP solution w.r.t. accurate neighborhood is utilized, the algorithm effectiveness is high in this case. However, the corresponding time cost increases as grouping takes no effect.

6. EXPERIMENTS

Experimental evaluation answers the following questions:

(1) *By utilizing dirty data, can it form more accurate clusters?* In Figure 6, we illustrate an example of artificial data points on how the dirty points affect the clustering and how the simultaneous clustering and repairing work. Figures 7, 8 and 11 compare quantitatively the clustering accuracy of our DORC to the existing DBSCAN (directly discarding all noisy data) on both synthetic and real data sets.

(2) *By simultaneous repairing and clustering, in practice is the repairing accuracy improved compared with the existing data repairing approaches?* Figures 9 and 12 compare our proposed DORC with the existing data repairing methods over real data sets.

(3) *How do the approaches scale?* In Figure 10, we show that it is possible to trade effectiveness for efficiency in LDORC via the group distance threshold τ. Figure 13 reports the scalability over a large-scale data set with up to 400k data points.

Our programs are implemented in Java and all experiments were performed on a PC with Intel(R) Core(TM) i7-2600 3.40GHz CPU and 8 GB memory.

Clustering Accuracy. To evaluate the clustering accuracy, we employ the purity measure [18]. It evaluates the most frequent class label of data points in each cluster,

$$\mathsf{purity} = \frac{1}{n} \sum_i \max_j |\mathsf{cluster}_i \cap \mathsf{class}_j|,$$

that is, counting the maximum number of data points in each cluster i corresponding to a class j. The higher the measure is, the better the clustering accuracy is.

Repairing Accuracy. The repair accuracy evaluates how close the repaired result $\lambda(p_i)$ is compared to the true location $\mathsf{truth}(p_i)$. We employ the repair error measure, root-mean-square error (RMS) [13],

$$\mathsf{error} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \delta(\mathsf{truth}(p_i), \lambda(p_i))^2}.$$

The lower the repair error is, the more accurate the repair is (closer to the ground truth).

6.1 Artificial Data Set

Example Results. To study the exact ILP and approximate DORC solutions, we draw a small synthetic dataset as shown in Figure 6. For the clean data in Figure 6(a), there are two classes, C1 and C2, with 164 data points. We introduce up to 34 artificial dirty points (by moving points in particular areas to random locations in a certain radius). Clustering approaches are performed over the data with dirty points.

Figure 6(b) presents the clustering results by DBSCAN over the dirty data (without repairing). Three clusters present and a number of 17 black points are identified as noises, i.e., not belonging to any cluster. Figure 6(c) reports the results by our proposed DORC of simultaneous repairing and clustering over the same dirty dataset. As illustrated in Figure 6(c), DORC can successfully repair the dirty points and return two clusters similar to the two classes in ground truth in Figure 6(a). These results verify our intuition that a large number of dirty points may greatly affect the clustering results (Figures 6(a) vs. 6(b)), while the clustering with repairing can address the variance introduced by dirty data (Figures 6(a) vs. 6(c)). The results illustrate that our proposal is not limited to splitting cluster (in Figure 1), but may also return merged clusters that are erroneously split by DBSCAN in Figure 6(b).

Quantitive Results. Figure 7 delivers the accuracies of clustering and repairing under various dirty rates over the synthetic dataset. A dirty rate 0.21 denotes that 21% points are modified as dirty data. Besides DBSCAN, the results of another density-based clustering, OPTICS [2], are also reported.

As shown in Figure 7(a), it is not surprising that the clustering accuracy of all approaches drops with the increase of

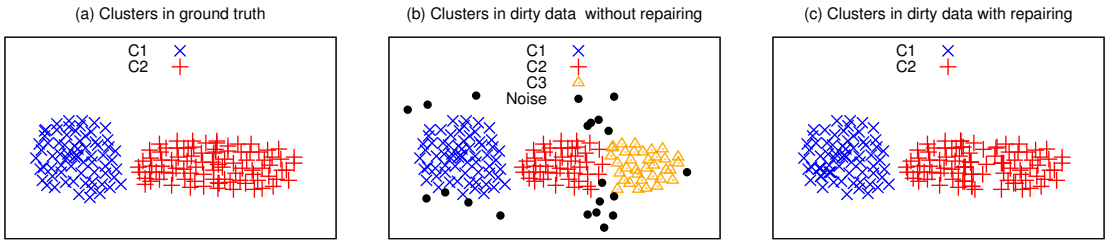


Figure 6: Clusters in (a) synthetic clean data, (b) dirty data without repairing, (c) dirty data with repairing

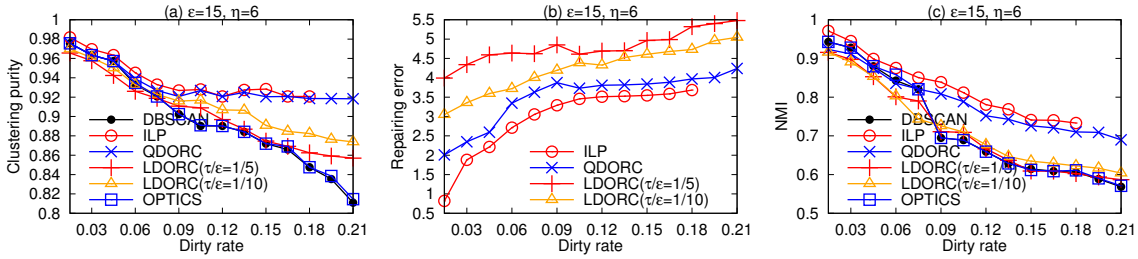


Figure 7: (a) Clustering accuracy, (b) repairing accuracy, and (c) NMI accuracy, over synthetic data

dirty rate. Most importantly, our DORC approaches can significantly improve the accuracy of clustering compared to DBSCAN, especially when the dirty rate is large. Moreover, the clustering purity of the approximation algorithm QDORC is comparable to that of the exact method ILP (we fail to obtain ILP results in dirty rates greater than 0.18 owing to the extremely high time costs).

Figure 7(b) reports repairing error. The exact method ILP has lower repairing error. The result verifies the rationale of considering the minimum cost repairs, following the minimum change principle in data repairing [24]. Nevertheless, the approximate QDORC approach is comparable with ILP, especially when the dirty rate is large. The reason behind is that with an extremely large amount of dirty data, the repairing (even the exact one) might not be precise.

6.2 Real GPS Data Set

For the real dataset, we collect[1] 818 GPS reading points in three nearby buildings, which correspond to three classes. Owing to the weak signal inside buildings, a large amount of GPS readings are inaccurate (139 points outside the buildings). We manually alter these (originally embedded rather than randomly introduced) dirty data points, i.e., label the true building for each inaccurate point as the ground truth. Clustering approaches are conducted over the dirty dataset, where the truth class of each dirty point is labeled. We use Euclidean distance as the distance function δ on GPS points.

Note that the GPS data are continuously collected in a time period. Filtering techniques can be applied to clean the noisy data in such a time-space correlated time-series [14], e.g., by the widely used Median Filter (MF) [22]. The main idea of MF is to go through the GPS readings one by one in the time-series, repairing each point with the median of (temporally) neighboring points. Therefore, instead of simultaneous repairing and clustering, MF+DBSCAN first applies MF to clean the GPS data, and then performs the existing DBSCAN clustering over the MF pre-processed data.

Figures 8 and 9 present the clustering and repairing accuracy results with various dirty rates, density thresholds η and distance thresholds ε (parameters η and ε are inherited from the density-based clustering DBSCAN, see [7] for a discussion on determining such parameters). The results are generally similar to that on the synthetic data in Figure 7. In addition, by applying MF, the clustering accuracy of DBSCAN is slightly improved. It verifies the motivation of this study, i.e., utilizing the dirty data can enhance clustering.

With various clustering requirements of ε and η in Figures 8 and 9, our proposed QDORC always shows aclear improvement. In particular, the clustering accuracy improvement by QDORC is more significant than that of MF+DBSCAN in Figures 8(b) and 8(c). The result is not surprising given the significantly lower repairing error of QDORC compared to MF in the corresponding Figures 9(b) and 9(c).

Figure 10 verifies the analysis at the end of Section 5 that τ in LDORC provides a trade-off in efficiency and effectiveness. As shown, a large $\tau = \frac{1}{2}\varepsilon$ shows high efficiency (low time cost) in Figure 10(c), while its clustering accuracy in Figure 10(a) is low and the repairing error in Figure 10(b) is high. On the other hand, a small τ leads to high time cost but lower repairing error and better clustering purity. When $\tau = 0$, the result of LDORC is exact the same as that of QDORC without grouping. The corresponding time cost (of $\tau = 0$) may be a bit higher since LDORC has extra cost on grouping.

6.3 Restaurant Data Set

Restaurant[2] is a collection of 864 restaurant records that contains 112 duplicates and is widely used for record matching [21]. Each group of duplicates can be interpreted as a

[1]Since existing datasets for evaluating clustering are not labeled for dirty data, we need to collect and manually label the ground truth of dirty data points (with a great manual effort).

[2]http://www.cs.utexas.edu/users/ml/riddle/data.html

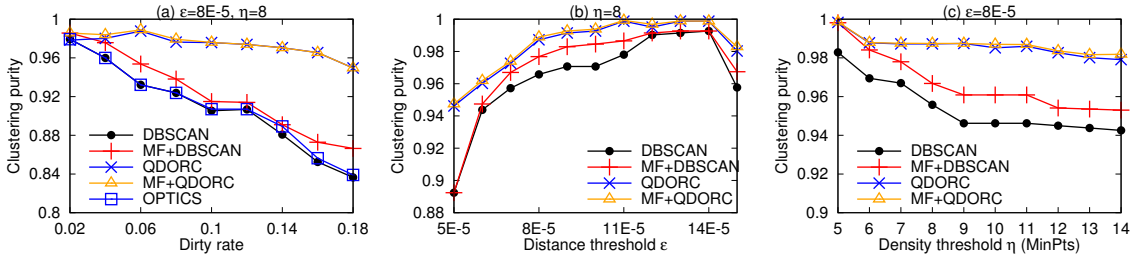


Figure 8: *Clustering accuracy* under various (a) dirty rates, (b) ε, and (c) η, over GPS data

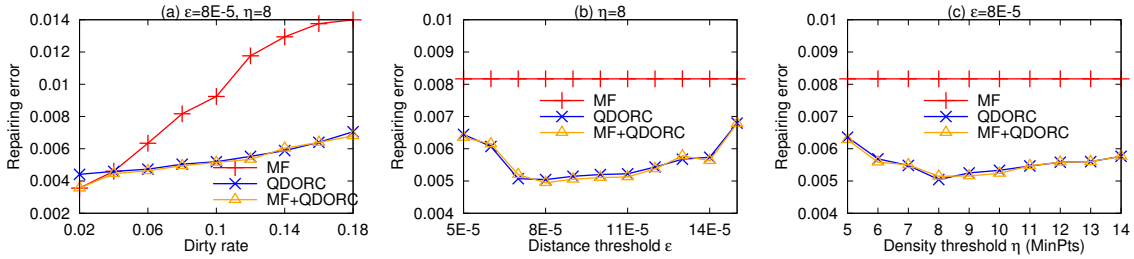


Figure 9: *Repairing accuracy* under various (a) dirty rates, (b) ε, and (c) η, over GPS data

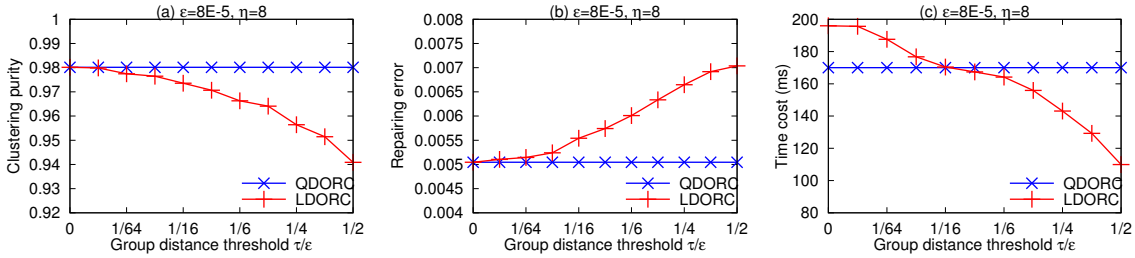


Figure 10: *Trade-off* in LDORC via group distance threshold τ over GPS data

class. A record includes four attributes, name, address, city and type. Edit distance [19] is employed as the distance function δ. Since the data are originally clean, following the same line of evaluating data repairing performance [3], we introduce dirty values by randomly replacing values (could be any data instead of one particular area) in the data set, with various dirty rates. Different from filtering numerical values on GPS data set, integrity constraints (such as FDs name,address $\rightarrow$ city) in databases are employed to clean the dirty data [3]. It is the reason *why we employ this data set*, where both the existing clustering and repairing techniques can be applied (as rational baselines).

The results in Figures 11 and 12 are generally similar to Figures 8 and 9 over GPS data. Figure 11 shows that by applying FD-based repairing first, the clustering accuracy is improved by FD+DBSCAN. Our proposed QDORC achieves significantly higher clustering and repairing accuracies. Similar results on various η and ε are also observed.

Most importantly, by combining our QDORC with the existing FD constraint-based repairing, i.e., FD+QDORC, both the clustering purity and repairing error are further improved compared to the FD(+DBSCAN) approach. The results demonstrate that the proposed DORC is *complementary to the existing constraint-based repairing*, by directly applying QDORC over the data repaired by FD.

6.4 Foursquare Data Set

The experiment on Foursquare dataset focuses on scalability over large data sizes, up to 400k check-in data points. Since this large scale data set is not pre-labeled, we mainly observe the time cost.

First, as shown in Figure 13(a), while the number of noise points increases as the data size, the number of leaders keeps low. It verifies the result in Proposition 9 that the number of leaders m is bounded. Consequently, the corresponding time cost of LDORC increases linearly, in Figure 13(b). The result verifies the linear time complexity of the LDORC algorithm in Proposition 6. Finally, Figure 13(c) reports a result similar to Figure 10(c) over GPS data that with the increase of τ, the efficiency is improved.

6.5 UCI Data Set

Finally, in order to show that the proposed approach improves the clustering accuracy on real data, we report experiments on two labeled publicly available benchmark data, Iris and Ecoli, from UCI[3]. Moreover, a state-of-the-art evaluation measure, normalized mutual information (NMI) [20], is employed for clustering validation. As shown in Figure 14, similar results are generally observed, i.e., our QDORC still

[3]http://archive.ics.uci.edu/ml/datasets.html

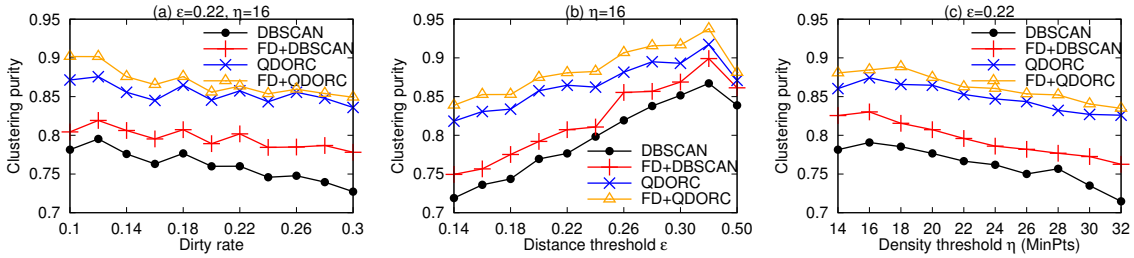


Figure 11: ***Clustering accuracy*** **under various (a) dirty rates, (b) ε, and (c) η, over Restaurant**

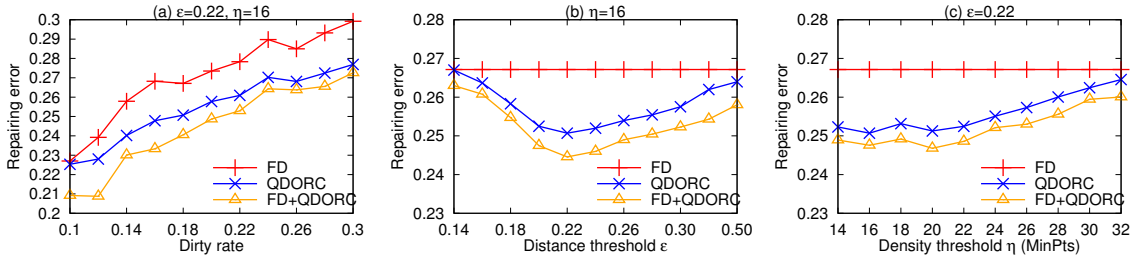


Figure 12: ***Repairing accuracy*** **under various (a) dirty rates, (b) ε, and (c) η, over Restaurant**

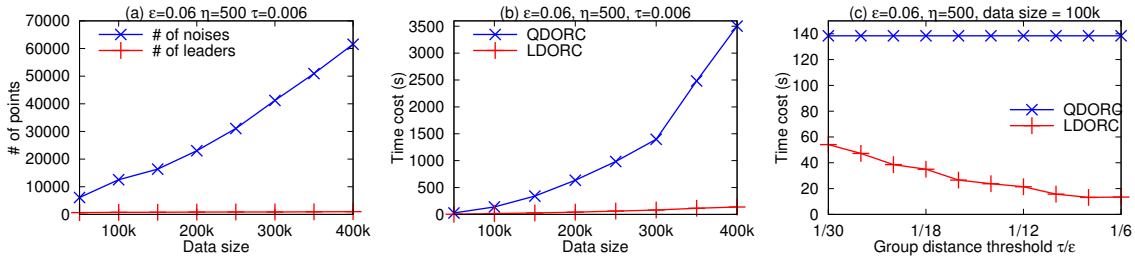


Figure 13: ***Scalability*** **over Foursquare data**

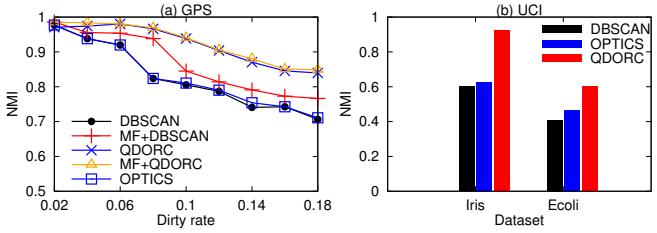


Figure 14: ***NMI clustering accuracy*** **on real datasets**

shows much higher (NMI) accuracy (compared to DBSCAN and OPTICS).

7. RELATED WORK

Clustering. Density-based clustering has proved to be useful in various applications (see [16] for a survey). Given a distance threshold Eps ε and a minimum requirement of neighborhood MinPts η, DBSCAN [7] determines the density region of core points as well as the corresponding border points. Efficient implementation [23] and incremental computation [6] of DBSCAN has been devised. In this study, we also follow the settings of Eps ε and MinPts η.

OPTICS [2] produces a cluster-ordering of data points with various Eps levels, and maintains a sorting of core points. In contrast, we rank non-core points in this paper, according to their likelihood of being core points after repairing.

DENCLUE [11] generalizes the notation of density clusters by introducing the concept of influence functions. The influence function models the influence of a point to its neighbors, e.g., by square wave functions or Gaussian functions. fDBSCAN [17] considers the extension over fuzzy/uncertain data points. Rather than pruning outliers, Gupta and Ghosh [9] identify dense regions of subsets of points. Xiong et al. [25] study clustering anomaly, where clusters are formed by a (very) small portion of points in a large data set. Again, all these studies focus on identifying noise/non-noise points, while the large amount of dirty data (identified as noises) are still not employed to form clusters. In contrast, our proposal considers cleaning and clustering *with* dirty data.

While outlier detection (see [12] for a survey) identifies dirty points, data repairing further modifies the points for correction. Indeed, our proposal incorporates density-based DBSCAN (that also identifies noises and could be regarded as an outlier detection method). In this sense, data repairing and outlier detection are complementary.

Repairing. Besides the minimum modification model [24], which is also adopted in our study, a deletion model [5] is often considered. The deletion model finds the minimum removal of dirty data. In this sense, DBSCAN is also a deletion-based cleaning technique that removes dirty data. As discussed, simply ignoring the large amount of dirty data as noises by the existing clustering approaches is not rational and may affect greatly the clustering results.

In addition to the widely used FD constraints, other types of data quality rules are employed (see [8] for a review). Again, as discussed in the introduction, our study considers only the density information embedded in data rather than the external knowledge of constraints or rules. Most importantly, following the same line of combining with FD-based repairing (in Section 6.3), our proposal is complementary to these state-of-the-art techniques, whenever extra information such as master data or constraint rules are available.

8. CONCLUSION

Preliminary density-based clustering can successfully identify noisy data but without cleaning them. On the other side, existing constraint-based repairing relies on external constraint knowledge without utilizing the density information embedded inside the data. In this paper, inspired by the aforesaid victory and defeat, we study a novel problem of clustering and repairing dirty data at the same time. To the best of our knowledge, this is the first study on enhancing clustering by repairing and utilizing dirty data. With the happy marriage of clustering and repairing advantages, *both the clustering and repairing accuracies are significantly improved* as presented in the experimental evaluation.

To tackle the DORC problem of simultaneous clustering and repairing, our major technique contributions include: (1) the formulation of DORC as an ILP problem; (2) an optimal solution to the corresponding LP relaxation that can be directly obtained without calling LP solvers; (3) a quadratic time approximation algorithm devised upon the LP solution; and (4) a linear time improvement via grouping data points. In particular, the linear time algorithm provides a trade-off between effectiveness and efficiency.

Acknowledgement

This work is supported in part by the Tsinghua University Initiative Scientific Research Program; China NSFC under Grants 61202008, 91218302 and 61370055; National Grand Fundamental Research 973 Program of China under Grant 2012-CB316200; Huawei Innovation Research Program.

9. REFERENCES

[1] Full Version. *http://ise.thss.tsinghua.edu.cn/sxsong/doc/cc.pdf.*

[2] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. In *SIGMOD Conference*, pages 49–60, 1999.

[3] P. Bohannon, M. Flaster, W. Fan, and R. Rastogi. A cost-based model and effective heuristic for repairing constraints by value modification. In *SIGMOD Conference*, pages 143–154, 2005.

[4] H. Chen, W.-S. Ku, H. Wang, and M.-T. Sun. Leveraging spatio-temporal redundancy for rfid data cleansing. In *SIGMOD Conference*, pages 51–62, 2010.

[5] J. Chomicki and J. Marcinkowski. Minimal-change integrity maintenance using tuple deletions. *Inf. Comput.*, 197(1-2):90–121, 2005.

[6] M. Ester, H.-P. Kriegel, J. Sander, M. Wimmer, and X. Xu. Incremental clustering for mining in a data warehousing environment. In *VLDB*, pages 323–333, 1998.

[7] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996.

[8] W. Fan. Dependencies revisited for improving data quality. In *PODS*, pages 159–170, 2008.

[9] G. Gupta and J. Ghosh. Robust one-class clustering using hybrid global and local search. In *ICML*, pages 273–280, 2005.

[10] M. A. Hernández and S. J. Stolfo. Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Min. Knowl. Discov.*, 2(1):9–37, 1998.

[11] A. Hinneburg and D. A. Keim. An efficient approach to clustering in large multimedia databases with noise. In *KDD*, pages 58–65, 1998.

[12] V. J. Hodge and J. Austin. A survey of outlier detection methodologies. *Artif. Intell. Rev.*, 22(2):85–126, 2004.

[13] S. R. Jeffery, M. N. Garofalakis, and M. J. Franklin. Adaptive cleaning for rfid data streams. In *VLDB*, pages 163–174, 2006.

[14] K. H. Ji and T. A. Herring. A method for detecting transient signals in gps position time-series: smoothing and principal component analysis. *Geophysical Journal International*, 193(1):171–186, 2013.

[15] S. Kolahi and L. V. S. Lakshmanan. On approximating optimum repairs for functional dependency violations. In *ICDT*, pages 53–62, 2009.

[16] H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek. Density-based clustering. *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, 1(3):231–240, 2011.

[17] H.-P. Kriegel and M. Pfeifle. Density-based clustering of uncertain data. In *KDD*, pages 672–677, 2005.

[18] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.

[19] G. Navarro. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88, 2001.

[20] X. V. Nguyen, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854, 2010.

[21] P. D. Ravikumar and W. W. Cohen. A hierarchical graphical model for record linkage. In *UAI*, pages 454–461, 2004.

[22] J. W. Tukey. Exploratory data analysis. 1977.

[23] P. Viswanath and V. Suresh Babu. Rough-dbscan: A fast hybrid density based clustering method for large data sets. *Pattern Recognition Letters*, 30(16):1477–1488, 2009.

[24] J. Wijsen. Database repairing using updates. *ACM Trans. Database Syst.*, 30(3):722–768, 2005.

[25] Y. Xiong, Y. Zhu, P. S. Yu, and J. Pei. Towards cohesive anomaly mining. In *AAAI*, 2013.