# CERTIFIABLE ROBUSTNESS FOR NEAREST NEIGHBOR CLASSIFIERS

Austen Z. Fan†, Paraschos Koutris†

†University of Wisconsin - Madison

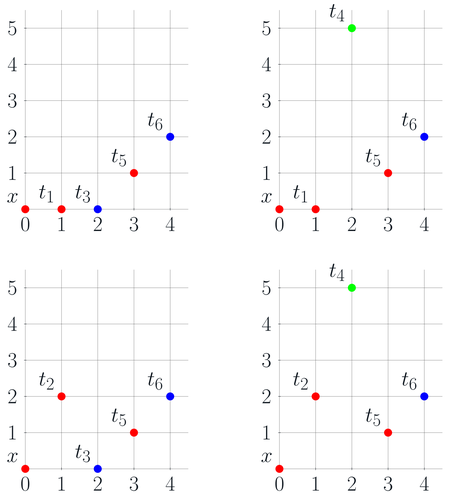
## A Motivating Question

One of the most useful features in DBMS is that one can specify and enforce **Functional Dependencies (FD)** on data. In practice, however, data stored might violate those pre-defined FDs due to various reasons, e.g. during data integration in a data warehouse [1]. Such database will be called **inconsistent**. We ask the following questions: How much should we trust a model prediction when the training data is inconsistent? In this paper, we study the notion of **Certifiable Robustness (CR)** as a measure of such confidence. Roughly speaking, a **Machine Learning (ML)** model is of CR if it classifies/predicts a test point consistently when trained on all possible **repairs** of the database. Here, a repair is a maximally consistent subset of the inconsistent database.

## An Example

| | A | B | C | Label |
|---|---|---|---|---|
| $t_1$ | 1 | 0 | a | 0 |
| $t_2$ | 1 | 2 | b | 0 |
| $t_3$ | 2 | 0 | a | 2 |
| $t_4$ | 2 | 5 | c | 1 |
| $t_5$ | 3 | 1 | a | 0 |
| $t_6$ | 4 | 2 | d | 2 |

Consider the above database where the FD specified is $A \to B$. Note that there are $2 \times 2 = 4$ possible repairs by choosing one of $t_1$ and $t_2$, and one of $t_3$ and $t_4$. We then visualize the repairs where colors refer to labels. Suppose we employ the $k$-Nearest Neighbor Classifiers to predict the label of a test point $x$ where $k = 3$, i.e. take the majority vote among the 3 nearest data points. Assume the distance metric is the 2-norm in attributes A and B, and $x$ corresponds to the point (0,0).




We observe that 3-NN will predict the test point to be label 0 (red) in all 4 repairs. We then say 3-NN is of CR with respect to the test point $x$ for our database. Label 0 is called the **certain label** for $x$.

## Problem Definition

We are thus interested in the following question. Given an inconsistent labeled instance $D$ over an FD schema $\mathbf{R}$ and a test point $x$, is $x$ certifiably robust for $k$-NN classification? We denote this problem as $\text{CR-NN}\langle \mathbf{R}, k \rangle$. We also consider its counting version, denoted as $\#\,\text{CR-NN}\langle \mathbf{R}, k, \ell \rangle$, which asks on how many repairs $k$-NN will classify $x$ to be the label $\ell$.

## Main Results

To state our main results, we need the notion of **left-hand-side chain (lhs chain)**. We say a set of FDs $\Sigma$ has a lhs chain if for every two FDs $X_1 \to Y_1$ and $X_2 \to Y_2$ in $\Sigma$, either $X_1 \subseteq X_2$ or $X_2 \subseteq X_1$ [2]. For example, the FD set $\{A \to C, B \to C\}$ does not have an lhs chain, while the FD set $\{AB \to C, B \to D\}$ has an lhs chain.
Let $\mathbf{R}$ be an FD schema. Our main result asserts the following dichotomy [3]:

- If $\mathbf{R}$ is equivalent to an FD schema with an lhs chain, then CR-NN$\langle \mathbf{R} \rangle$ and $\#$CR-NN$\langle \mathbf{R} \rangle$ can be computed in polynomial time.
- Otherwise, for any integer $k \geq 1$, CR-NN$\langle \mathbf{R}, k \rangle$ is coNP-complete and $\#$CR-NN$\langle \mathbf{R} \rangle$ is #P-complete.

## Algorithms

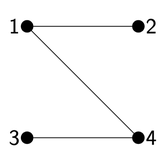
We design a polynomial time algorithm when FD schema is in the tractable case. The algorithm is inspired by the OptSRepair algorithm in [4]. When the FD schema is given by a single primary key, we design a linear-time algorithm which vastly generalizes the SortScan (SS) and the MinMax (MM) algorithms in [5]. We make a comparison in the following table, where $|D|$ is the size of the inconsistent database and $m$ is the number of possible labels.

| | Our algorithm | Karlas et al.'s |
|---|---|---|
| Time Complexity | $O(\lvert D\rvert \cdot m)$ | $O(\lvert D\rvert \cdot m)$ for MM<br>$\Omega(\lvert D\rvert \cdot \binom{m+k-1}{k})$ for SS |
| Applicability | any $k$ and $m$ | MM only for $m = 2$<br>only under a further restriction on labels |

We note that the general algorithm runs in $O(n^c)$ for some constant $c$ depending on the structure of the FD schema. This may hinder practical implementation when $c$ is not small. See more discussions on this issue under the Open Problems.

## A Taste of Hardness

The proof of hardness for decision problem $\text{CR-NN}\langle \mathbf{R}, k \rangle$ is a bit involved. That said, we offer a taste of the hardness proof. One important step is to view a maximal matching of a bipartite graph $G$ as a repair of a labeled instance $D$ with FD schema $\{A \to B, B \to A\}$. This can be seen from the following example. Given a graph, we can associate it with the database listed on its right side, as illustrated below.




FD: A → B, B → A

| A | B |
|---|---|
| 1 | 2 |
| 1 | 4 |
| 3 | 4 |

Therefore, a maximal matching of the graph corresponds to exactly one repair of the database and vice versa. This can be generalized easily to arbitrary bipartite graphs.

## MIN-REPAIR and FORBIDDEN-REPAIR

We consider two variants of OPT-REPAIR [4], called MIN-REPAIR and FORBIDDEN-REPAIR, and relate them to the $\text{CR-NN}\langle \mathbf{R}, k \rangle$ problem. MIN-REPAIR asks one to find the subset repair that has tuples with the *smallest total weight*. In FORBIDDEN-REPAIR, one is given an inconsistent instance $D$ and a subinstance $S \subseteq D$, and is asked to find a subset repair $I \subseteq D$ such that $I \cap S = \emptyset$.
Note that MIN-REPAIR captures as a special case of FORBIDDEN-REPAIR. We show that there exists a many-one polynomial time reduction from FORBIDDEN-REPAIR to the complement of CR-NN$_<\langle \mathbf{R}, 1 \rangle$.

## CR Under Other Uncertain Models

We also consider CR for three widely used uncertain models.
**?-Sets with Size Constraints** [6] For a given instance $D$ over the schema, we mark an uncertain subset $D_?$ of the tuples in $D$. Then, for a positive integer $m \geq 1$, we define the set of possible repairs as: $\mathcal{I}_? = \{I \,|\, D \backslash D_? \subseteq I \subseteq D, |\, D \backslash I \,| \leq m\}$.
**Or-Sets** [7, 8] In this uncertain model, each attribute value of a tuple is an or-set consisting of finite values. Each possible repair in $\mathcal{I}_{\text{or}}$ is formed by choosing exactly one value from each or-set, independent of the choices across all other or-sets.
**Codd tables** [5] In a Codd table, a missing value is represented as Null paired with a domain from which that value can be chosen. A repair is any possible completion of the table.
We give $\text{CR-NN}\langle \mathbf{R}, k \rangle$ polynomial time algorithms for all three uncertain models (the algorithm for ?-sets with size constraints and Codd tables run in linear time).

## Open Problems

Many interesting questions remain open at this point. For example, on the theory side, one can consider CR

- for other widely used classification algorithms, such as decision trees, Naive Bayes classifiers and linear classifiers.
- when the instance $D$ is not a single table but the join of several tables.
- for other integrity constraints, e.g. inclusion dependencies.

On the practical side, one can consider improving the current result by

- deriving fast heuristic or approximation algorithm for $\text{CR-NN}\langle \mathbf{R}, k \rangle$ or $\#\,\text{CR-NN}\langle \mathbf{R}, k \rangle$ (under some assumption on simple FD schema structure).
- deriving fast heuristic or approximation algorithm for (almost uniformly) sampling a repair that predicts a certain label.

## Acknowledgements

This research was supported in part by National Science Foundation grants CRII-1850348 and III-1910014, as well as a gift by Google.

## References

[1] Lukasz Golab. "Data Warehouse Quality: Summary and Outlook". In: *Handbook of Data Quality*. Springer, 2013, pp. 121–140.

[2] Ester Livshits, Benny Kimelfeld, and Jef Wijsen. "Counting subset repairs with functional dependencies". In: *J. Comput. Syst. Sci.* 117 (2021), pp. 154–164.

[3] Austen Z. Fan and Paraschos Koutris. "Certifiable Robustness for Nearest Neighbor Classifiers". In: *ICDT*. Vol. 220. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022, 6:1–6:20.

[4] Ester Livshits, Benny Kimelfeld, and Sudeepa Roy. "Computing Optimal Repairs for Functional Dependencies". In: *ACM Trans. Database Syst.* 45.1 (2020), 4:1–4:46.

[5] Bojan Karlas et al. "Nearest Neighbor Classifiers over Incomplete Information: From Certain Answers to Certain Predictions". In: *Proc. VLDB Endow.* 14.3 (2020), pp. 255–267.

[6] Anish Das Sarma et al. "Working Models for Uncertain Data". In: *ICDE*. IEEE Computer Society, 2006, p. 7. DOI: 10.1109/ICDE.2006.174.

[7] Samuel Drews, Aws Albarghouthi, and Loris D'Antoni. "Proving data-poisoning robustness in decision trees". In: *PLDI*. ACM, 2020, pp. 1083–1097.

[8] Sergio Greco, Cristian Molinaro, and Francesca Spezzano. *Incomplete Data and Data Dependencies in Relational Databases*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2012.