Reputation-Aware Data Fusion and Malicious Participant Detection in Mobile Crowdsensing

Yujian 'Charles' Tang*+, Samia Tasnim*, Niki Pissinou, S. S. Iyengar, Abdur Shahid
School of Computing and Information Sciences
Florida International University
Miami, Florida 33199
Email: {stasn002, pissinou, ashah044}@fiu.edu, iyengar@cis.fiu.edu
+University of North Carolina, Chapel Hill, NC 27599
Email: {yujian@live.unc.edu}
* Both authors contributed equally to this work

***Abstract*—Mobile crowdsensing, an emerging sensing paradigm, promotes scalability and reduction in the deployment of specialized sensing devices for large-scale data collection in a decentralized fashion. However, its open structure allows malicious entities to interrupt a system by reporting fabricated or erroneous data, making trust evaluation a highly important issue in mobile crowdsensing applications. The goal of this research is to show that an introduction of a reputation system in the process of correlated sensor-based data fusion will enhance the overall quality of the sensed data. To do so, we design a reputation-aware data fusion mechanism to ensure data integrity. We use Gompertz function in our reputation method to rate the trustworthiness of the data reported by a crowdsensing participant. The proposed mechanism, on one hand, is capable of defending a data corruption attack and identifying malicious and honest participants based on their reported data in real-time. On the other hand, this mechanism yields more accurate data prediction in terms of lower data prediction error. We conducted experiments using two different real-world datasets. We compare our correlated data and reputation-aware data prediction (CDR) method with other popular methods, and the results show that our effective method incurs lower data prediction error.**

***Key words*- Big data analytics; anomaly detection; data fusion; mobile crowdsensing; Spatial-temporal data analysis**

I. Introduction

With the advent of better wireless technology and an increase in smartphone usage, a new mode of data collection named mobile crowdsensing (MCS) has emerged. Mobile crowdsensing has a number of practical applications: traffic monitoring, epidemic disease monitoring, reporting from disaster situations and environment monitoring [1], [2], [3]. For example, an environmental air quality sensing system was deployed on street sweeping vehicles to monitor air quality in San Francisco [4]. These applications are usually open to the public and receive sensor data from multiple participants. This influences the reduction of data sparsity at lower costs in comparison with traditional sensor networks. With various advantages, MCS's people-centric architecture allows both more inaccurate and corrupted data [5]. Malicious participants can manipulate the MCS data collection process at ease. These entities can interrupt a system by reporting fabricated or erroneous data, making trust evaluation a highly important issue in MCS applications. Therefore, validating the accuracy of contributions is essential to ensure the reliability of the application system.

In this paper, we consider data corruption attack behavior of a malicious participant. By malicious we mean a participant who sends incorrect data either intentionally or unintentionally. The unintentional error can arise because a participant carelessly performed the sensing task, or due to a sensor error. On the contrary, a malicious participant can deliberately fabricate the sensed data to infiltrate the system. For example, in air quality monitoring, a malicious participant may hold the sensor beside a burning cigarette or place it over sand instead of facing to the air. Thus, the reported data will not represent the actual air quality. In the related contemporary works [6], [7], [8], [9], the authors did not consider the participants' malicious behavior. Thus, these works were not able to distinguish the sensing data reported by malicious or careless users. This limitation of the existing works motivates us to design reputation-aware real-time data fusion algorithms for MCS to ensure data integrity. Our method can detect malicious participants and prevent them from infiltrating the system in real time.

We develop an online method for data quality prediction in MCS considering the heterogeneous trust level of the participants. We took into account spatial-temporal and inter sensor-category correlations. We consider the users who are willing to participate in sensing at the same time. The terms $participant$ or $node$ are used to denote a user with sensing capability.

We implement our Correlated Data and Reputation-aware Data Prediction (CDR) method on two real-world datasets [10], [11]. The sensing was performed for four days, and there are 289 taxi values in the first real dataset. The taxis move around different parts of Rome sensing temperature. The second data set consists of Beijing's air quality data. One hundred and forty nine taxis with four types of sensors collect $PM2.5, PM1.0, NO_2$ and $humidity$ data from Beijing during seven days.

The main contributions of this paper are as follows.

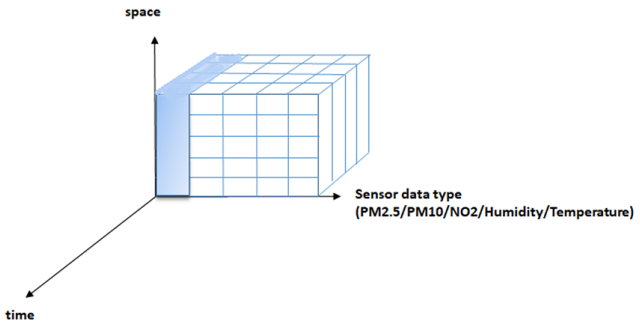

Fig. 1. Three-dimensional Tensor

• We propose a novel reputation-aware correlated sensor-based light-weight real-time data fusion and malicious participant detection mechanism for mobile crowdsensing data streams.

• Extensive experiments using two real-world data sets ensure the efficiency and accuracy of our data prediction mechanism over state-of-the art techniques.

We organize our paper as follows. In section II, we discuss the related work. In section III, we discuss different modules of our overall system. In section IV, we discuss our performance evaluation. Finally, concluding remarks and future works are offered in section V.

II. Related Work

In this section, we discuss the most pertinent works. Zhang et al. [12] proposed data cleaning method for environmental sensing which was based on incrementally adjusted reliability of individual sensors. With the advance of time, they incrementally adjusted the reliability of each sensor depending on the sensing data accuracy. Trustworthiness has been considered as a measure of data quality estimation [13]. Huang et al. [14] showed that using a reputation framework helped to weed out non-colluding malicious attackers. Their reputation framework produced more accurate results than not using a reputation framework. However, the authors assumed that data is coming from every discrete block of space-time which is not practical in real-world scenarios. On the contrary, Peng et al. [15] used unsupervised learning for data quality estimation. Though this method works after the collection of historical data from all the users, it is not an online method.

Nowadays, instead of traditional static wireless sensor networks, the sensing is distributed among a crowd of people. This brings heterogeneity in the sensor networks and makes the computation more complex. The most recent work on data quality estimation in mobile crowdsensing is done by Shengzhong et al. [16]. The authors broached real time data estimation in mobile crowd sensing and proposed a context-aware method for data quality estimation. The limitation of this work is that the authors considered the presence of exactly one mobile user at each point of interest (PoI). Kishino et al. [17] mounted sensor nodes on garbage trucks that drive around the city. Their motivation was to detect target events by analyzing vehicle-mounted sensor data streams. The authors used machine learning methods to achieve so. On the other hand, the author [7] broached a new sampling method named stratified sampling for calculating mean temperature of a linear area. In this paper, only random waypoint mobility model has been considered for the movement of the sensing devices.

There are several works focusing on the cleaning of data streams. Most of the previous works on sensor data cleaning focused on the reduction of consumed energy. To achieve this reduction, the authors [18], [19], [20] tried to reduce the inter node communication. In these works, it was assumed that sensor data are always aggregated during submission. There have been significant works on using compressive sensing for data reconstruction in static sensor networks [21], [22]. In recent days, researchers [23], [24], [25], [26], [16], [27] are designing frameworks to deal with big data services. In the past, the data size was not as big as present days, which influences researchers to design and develop scalable mechanisms to correct any kind of inaccuracy in data streams. Liu et al. [26] designed a framework for big data cleaning. This paper gives direction on how to achieve reliable database in big data applications. They used context to find similarity between data items. Moreover, the authors exploited usage pattern to classify and group data items that are not related contextually. One of the challenging tasks in dealing with big data is to shrink the data size by extracting the irrelevant subset. Dong et al. [28], in contrast, debated that having more data does not always provide more information. During data integration, proper selection of reliable source among all available sources results in higher data accuracy.

Another aspect of literature focuses on finding outliers in sensor data streams. In order to find global outliers in the data, Branch et al. [18] proposed a distance based ranking method. The other existing methods for finding outliers in sensor data are geometry-based [29] , polygon-based spatial outlier detection [30], clustering-based [31], kernel density-based [8] and histogram approach [32]. Bosman et al. [33] tried to answer the question if adding more neighbors makes the anomaly detection perform better. This paper considered static sensor nodes and it varied the neighborhood size by changing the communication range of the sensors.

However, to the best of our knowledge, we are the first to develop reputation-aware correlated sensor-based real-time data fusion and malicious participant detection mechanism for mobile crowdsensing data streams.

III. Methodology

In this section we first present an overview of the proposed mechanism, Correlated Data and Reputation-aware Data Prediction (CDR), then a detailed description of the components, and finally how we fit them together to create our full structure.

A. Overview

CDR consists of two parts: a reputation calculation method and correlated data [6]. The reputation method considers two types of trust for each sensor, *cooperation* and *reputation*, and both parameters are calculated at the application server level. The reputation calculation method is applied to multiple types of sensor data streams. These varied sensors are correlated with each other. It is important for our mechanism to take the granularity of time and space into account. We discretized our time into epochs, and space into equal-sized grids. The framework is applied only on data from sensors within the same *region* and the same epoch. CDR is applied to each different *type* of data and then the final, discretized space-time blocks are used to produce a least-square regression on the target data type. This regression can be used to predict both future data and missing data. We borrow the concept of three-dimensional tensors shown in Fig. 1 from [6]. The authors considered temporal interpolation for the sparse regions. However, Kang et al. [6] assumed that all incoming data from sensors was accurate.

B. Cooperation

Cooperation scores of sensors are measured per epoch; they measure the proportion of the inverse square root error of the data from the sensor over the sum of the proportion of the inverse square root error from all sensors. For our cooperation parameter, we used an inverse proportion of the square root of the absolute error so as not to punish small deviations from the average as much. In the data sets we tested, temperature data and air quality data, small variations from the average are common. The equation for cooperation score is shown in Eq. 1.

$$p_i = \frac{\frac{1}{\frac{\sqrt{|x_i - r|}}{\sum_{i=1}^{n}\sqrt{|x_i - r|}}}}{\sum_{j=1}^{n}\frac{1}{\frac{\sqrt{|x_j - r|}}{\sum_{i=1}^{n}\sqrt{|x_i - r|}}}} \tag{1}$$

Where r is the *robust average* of the data in that epoch and x_i is the measurement from sensor i. The *robust average* of the data provides an idea of where the data clusters, and this increases the accuracy of the data by assigning more weight to values that occur more frequently. We calculate *robust average* using Eq. 2.

$$r = \sum_{i=1}^{n} p_i * x_i \tag{2}$$

C. Reputation

Reputation scores are updated at the end of each epoch; it measures how accurate the crowdsensing participant has been over time. To calculate reputation from cooperation scores, first the cooperation scores are normalized [14] using Eq. 3. Here, P_i is the cooperation score of participant i. $min(p)$ and $max(p)$ denote the minimum and maximum cooperation score among all the participants during that epoch. After normalization, the cooperation scores belong to the range $[-1, 1]$.

$$p_i^{norm} = \frac{2(p_i - min(p))}{max(p) - min(p)} - 1 \tag{3}$$

We want to maximize impact of the most recent epochs and minimize the impact of the least recent ones. To make the aging effective, we *age* the normalized cooperation scores with Eq. 4.

$$p_{i,k}^{'} = \sum_{k'}^{k} \lambda^{k-k'} p_{i,k}^{norm} \tag{4}$$

Here, k denotes the current epoch and k' has the value from 1 to current epoch. Aging parameter λ has the value $[0, 1]$ Finally, reputation is calculated using the Gompertz function [14], shown in Eq. 5.

$$R_{i,k} = ae^{be^{cp_{i,k}^{'}}} \tag{5}$$

Here, a, b and c are function parameters. The parameter a denotes the upper asymptote, displacement along x-axis is controlled by b and the growth rate is controlled by parameter c.

D. Full Structure

We discretize the space into *regions* and the time into epochs, then we run CDR on every discrete block of space-time.

First we run an Expectation Maximization Algorithm (EM), shown in Algorithm 1, on the "reputable" sensors. To be classified as reputable sensors, the participant must have a reputation higher than the threshold. This threshold is an application dependent. Initially, all sensors are classified as reputable with equal cooperation score.

Algorithm 1 Expectation Maximization on Cooperation Scores for Robust Average

Input: Robust Average (r) , Cooperation Scores (p_i)
Output: Robust Average (r)
Initialize: all p_i to $1/n$, where n is the number of sensors, and $l = 0$, where l is the iteration
while p_i^l *and* p_i^{l+1} *don't converge* **do**
  Compute r^{l+1} from p_i^l's using Eq. 2
  Compute p_i^{l+1}'s from r^l using Eq. 1
  $l = l + 1$
end
return r^{l+1}

After running EM algorithm once on only the reputable sensors, we then check the reported values from "disreputable" sensors, or sensors with a reputation lower than the threshold. If the reported value from any of these sensors is within an acceptable error range of the *robust average* calculated from the reputable sensors' reported data, then it is added as faux reputable sensor in that block of space-time. After

finding all the sensors from the set of disreputable sensors that contributed acceptable data in the block of space-time, EM is then run again on the new set of reputable sensors. The reason that we run EM twice is to provide sensors in the disreputable set a chance to move into the reputable set if they consistently contribute accurate data, because only sensors with a cooperation score for the epoch will have their reputations updated. The second EM run gives a new reputable average as well as update reputation scores for each sensor.

The new reputation scores are then normalized to the range $[-1, 1]$ using Eq. 3. The normalized cooperation scores are then aged based on their cooperation rating. Sensors with a cooperation score above a certain threshold are labeled as "cooperative" and sensors with a cooperation score below that threshold are labeled as "uncooperative". Depending on the sensor's classification for the latest block, the normalized cooperation is multiplied by a different aging parameter, λ. Cooperative sensors are multiplied by a lower aging parameter than uncooperative sensors. This means that the growth and decay rates of reputation will be different; the decay rate will be higher, and this provides higher punishment for bad data and thus helps quickly detect malicious users. Finally the aged cooperation score is inputted to Eq.5.

Once all the blocks are processed for each data *type*, then we use the processed data to create a least-square fit with the non-target data as the coefficient matrix, A, and the target data type as the dependent matrix, b as shown in Eq. 6.

$$\mathbf{A}\hat{x} = \mathbf{b} \tag{6}$$

The regression, $\hat{x}$, is then used to predict the target value given knowledge of all the other data values.

IV. Performance Evaluation

We used percentage absolute difference and Root Mean Square Error ($RMSE$) as performance metrics of data prediction accuracy. We compared the performance of our CDR method against mean-based and temporal linear regression-based data prediction models. We tested using two real-world data sets. In the first data set, our target type is temperature and uses two types of simulated correlated data. In the second data set, our target type is particulate matter with a diameter under 2.5 μm ($PM2.5$) and uses three types of real correlated data ($PM1.0, NO_2$ and $humidity$).

A. Temperature

The temperature data was from an area of roughly $22km$ by $23km$ and was taken over four days. The experimental area was split into 25 *regions* using a 5x5 equal-sized grid. We split the execution time into 96 epochs with each epoch being one hour long. We tested the performance of our CDR method against the existing mean-based method in three test data sets. To imitate the data impurity, continuous or random errors were applied on the temperature data streams. The data error from malicious participants ranged from 25% to 75%. Figures 2 through 4 show CDR's percentage improvement over the mean-based method, and each figure shows 612 predictions.

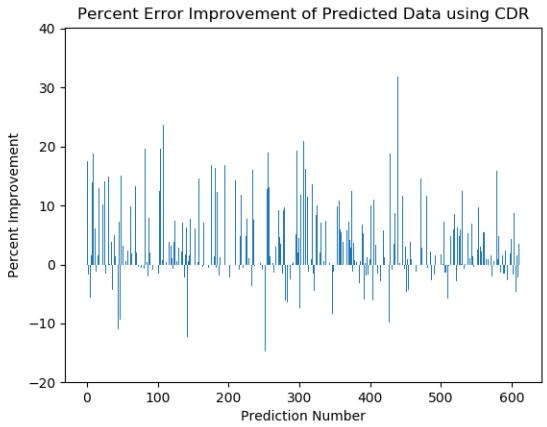

Fig. 2. Prediction results for test set 1: out of 612 predictions, CDR performed better in 466 and was within 5% of the true value in 290 cases

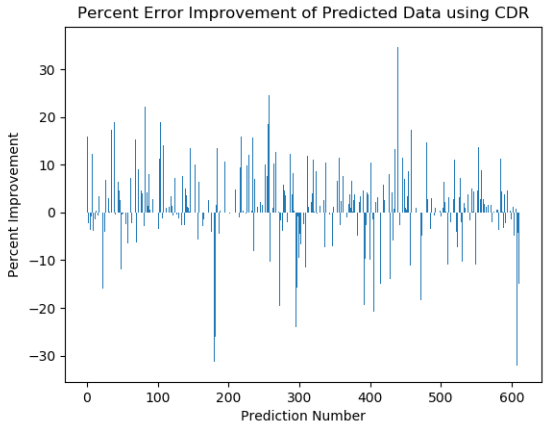

Fig. 3. Prediction results for test set 2: out of 612 predictions, CDR performed better in 453 and was within 5% of the true value in 261 cases

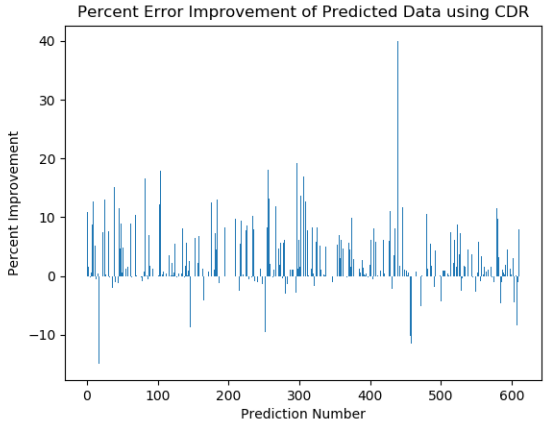

Fig. 4. Prediction results for test set 3: out of 612 predictions, CDR performed better in 498 and was within 5% of the true value in 213 cases

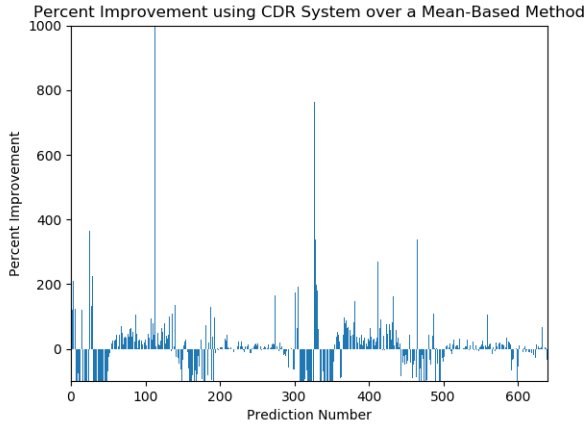

Fig. 5. Prediction results for test set 1: out of 640 predictions, CDR performed better in 379 cases

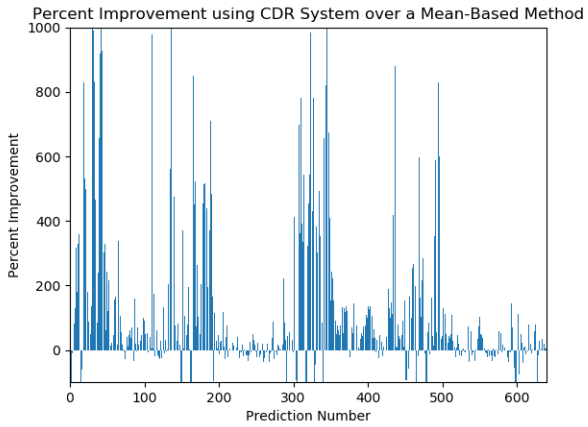

Fig. 6. Prediction results for test set 2: out of 640 predictions, CDR performed better in 445 cases

On average CDR was 16% more accurate and performed better in 77 percent of cases. Our CDR method incurred a cumulative percentage error of 9.3%.

B. PM2.5

The air quality data was collected from an area of roughly $120km$ by $150km$. The duration was seven days (149 hours). CDR was tested against the existing mean-based and temporal linear regression-based data prediction methods on five test data sets. To imitate the data impurity, continuous or random errors were applied on the crowdsensing data streams. The data error ranged from 25% to 75%.

We tested the performance of our algorithm for different levels of erroneous data from malicious users. We also varied the knowledge level of the participants in regards to the experimental environment to imitate sophisticated data manipulation by a malicious crowdsensing participant. Test set 1 (Fig. 5, Fig. 10, Fig. 15) was used for missing data prediction. We tested with sequential and random data loss patterns. In the first experiment with erroneous data from malicious users (Fig. 6, Fig. 8, Fig. 11, Fig. 13, Fig. 16, Fig. 18), we assumed the participants did not have any prior knowledge about the experimental environment. The data error ranged from 25% to 75%. One group of malicious participants reported a fixed percentage of error throughout the experiment. In the second experiment, we considered that the malicious participant has extended knowledge about the sensing area (Fig. 7, Fig. 9, Fig. 12, Fig. 14, Fig. 17, Fig. 19). Thus, these participants try to change the sensing data by adding noise to the air quality data of that particular spatio-temporal unit.

1) Percent Error per Prediction: Figures 5 through 9 show CDR's percentage improvement over the mean-based method, and each figure shows 640 predictions. On average CDR performed better in 70% of cases and is 70% more accurate.

2) Root Mean Square Error by Epoch: Figures 10 through 19 show CDR's improvement of the root mean square error ($RMSE$) normalized by epoch. We calculated $RMSE$ and

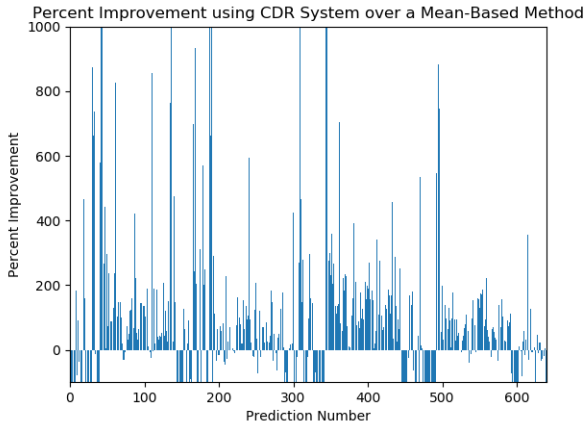

Fig. 7. Prediction results for test set 3: out of 640 predictions, CDR performed better in 442 cases

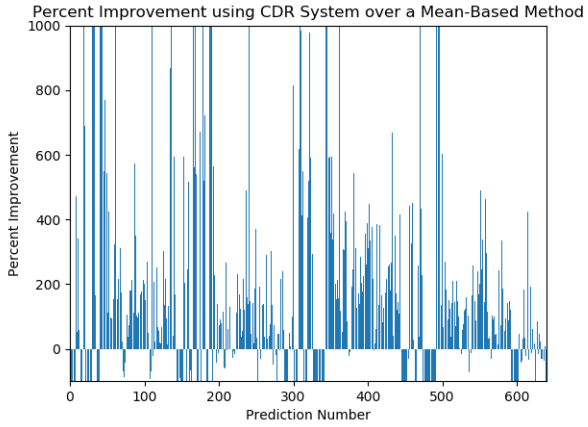

Fig. 8. Prediction results for test set 4: out of 640 predictions, CDR performed better in 454 cases

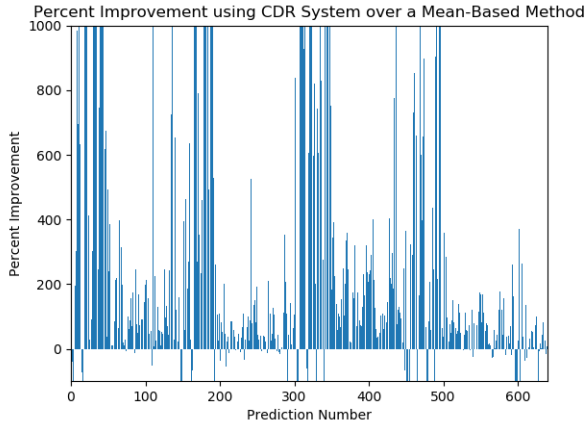

Fig. 9. Prediction results for test set 5: out of 640 predictions, CDR performed better in 533 cases

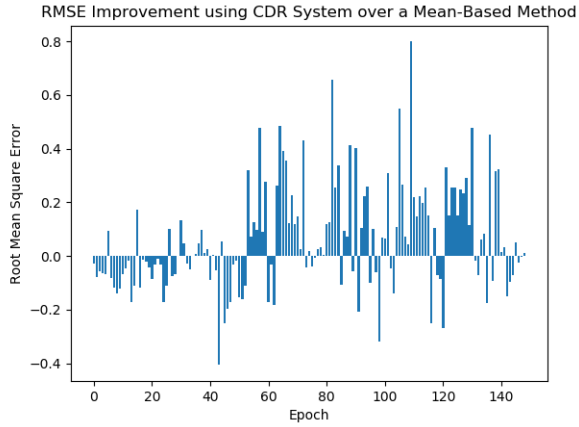

Fig. 10. Prediction results for test set 1: out of 149 epochs, CDR performed better in 88 epochs

used it as a performance measurement criteria of our algorithm. $RMSE$ is a standard metric to evaluate the accuracy of the prediction model [12].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{V}_i - V_i)^2}, \tag{7}$$

where $\widehat{V}_i$ is the predicted value, V_i is the original value and n is the number of epochs.

On average CDR had a lower $RMSE$ than mean-based method in 64 percent of the epochs and had a lower $RMSE$ by 25% . CDR's average $RMSE$ was 0.66, the average value of the target data type, $PM2.5$, was 79 with a range of $[4, 244]$.

Figures 15 through 19 show CDR's improvement in $RMSE$ over a temporal linear regression-based data prediction model. On average CDR incurred a lower $RMSE$ than the linear regression model by 59%, and performed better in 71 percent of epochs.

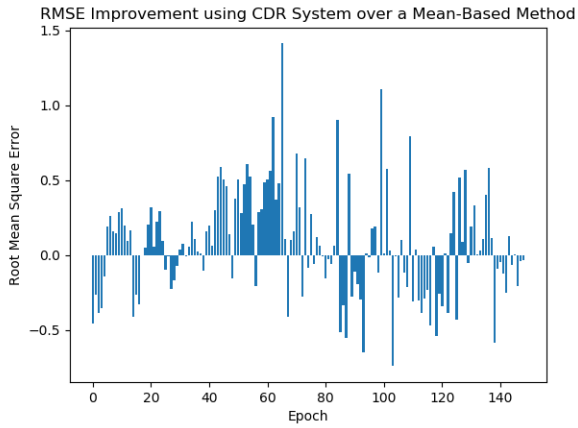

Fig. 11. Prediction results for test set 2: out of 149 epochs, CDR performed better in 88 epochs

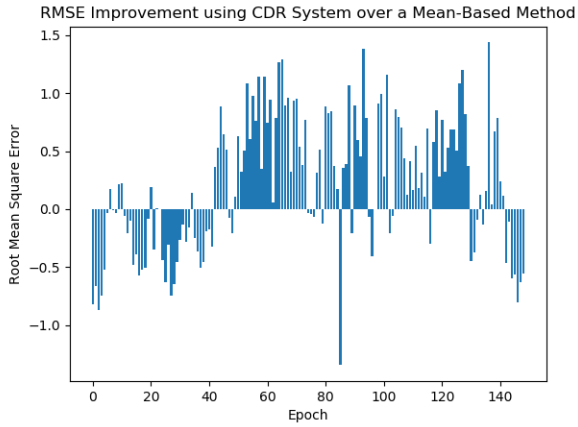

Fig. 12. Prediction results for test set 3: Out of 149 epochs, CDR performed better in 90 epochs

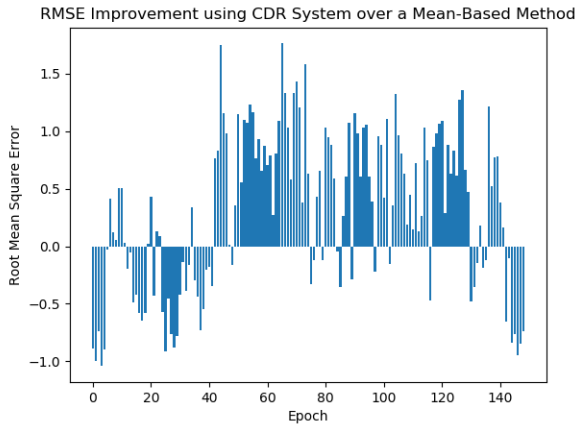

Fig. 13. Prediction results for test set 4: Out of 149 epochs, CDR performed better in 96 epochs

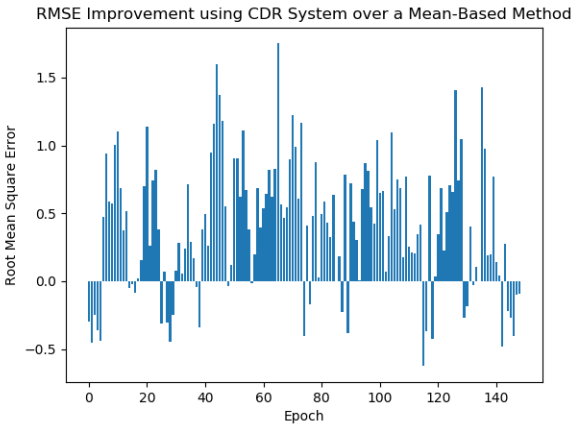

Fig. 14. Prediction results for test set 5: out of 149 epochs, CDR performed better in 115 epochs

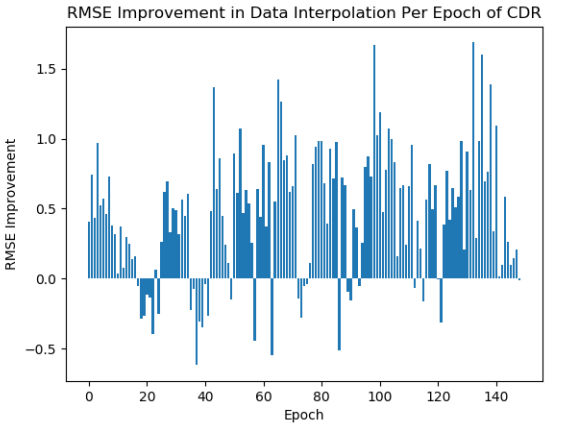

Fig. 15. Prediction results for test set 1: out of 149 epochs, CDR performed better in 119 epochs

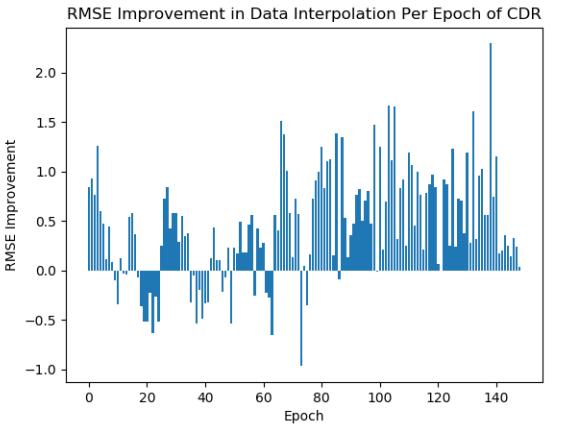

Fig. 16. Prediction results for test set 2: out of 149 epochs, CDR performed better in 119 epochs

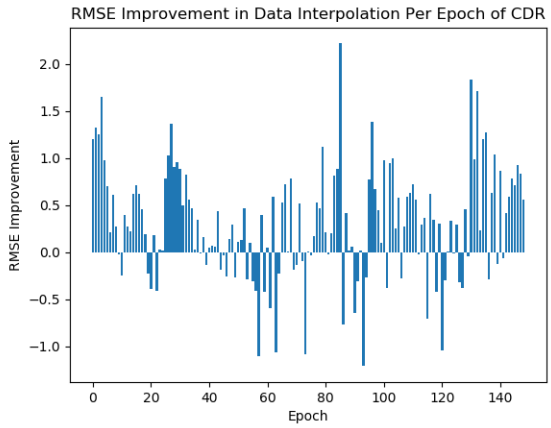

Fig. 17. Prediction results for test set 3: out of 149 epochs, CDR performed better in 105 epochs

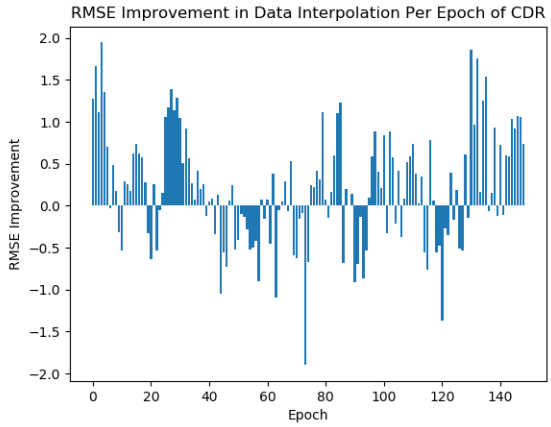

Fig. 18. Prediction results for test set 4: out of 149 epochs, CDR performed better in 93 epochs

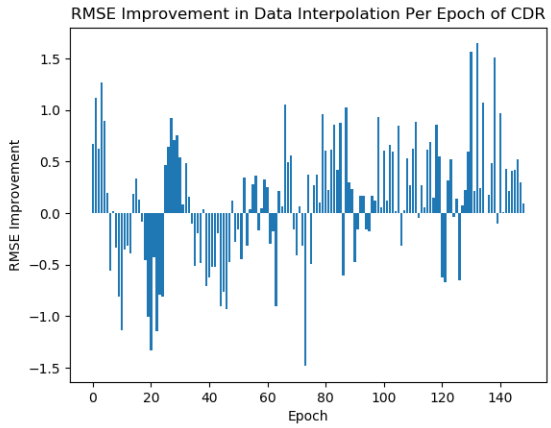

Fig. 19. Prediction results for test set 5: Out of 149 epochs, CDR performed better in 95 epochs

V. Conclusion & Future Work

In this paper, we proposed a novel method, named as CDR, for reputation-aware data fusion for mobile crowdsensing data streams. We showed that the proposed mechanism outperforms the existing mean-based and temporal linear regression-based data prediction models. We evaluate the approaches based on two datasets: Rome crowdsensing temperature and Beijing Air quality datasets, to demonstrate CDR's efficacy in different scenarios. For the Rome crowdsensing dataset, we achieved 16% better accuracy. Specifically, the 9.3% prediction error in temperature measurements of our approach equates to roughly 1 degree difference, which is negligible in real-life applications. With this in mind, we can say that our mechanism predicts temperature values with high accuracy. In case of the air quality dataset, our CDR method incurred on average 25% and 59% less $RMSE$ than mean-based and temporal linear regression models, respectively. Our data fusion method incurred an average $RMSE$ of 0.66 per epoch, which insinuates higher data prediction accuracy. The success of our approach lies in the integration of dynamic trust evaluation of the sensed data which allows us to defend data corruption attack and identify malicious and honest participants based on their reported data in real-time. In the future, we will extend this work considering collusion attack of malicious participants.

Acknowledgment

The work was supported by the National Science Foundation Grant number CNS-1560134 for the Research Experience for Undergraduates, Advanced Secured Sensor Enabling Technologies, and the Dissertation Year Fellowship support provided by Florida International University's Graduate School. The authors would like to thank Eric Xu for his contribution.

References

[1] F. Chen, P. Deng, J. Wan, D. Zhang, A. V. Vasilakos, and X. Rong, "Data mining for the internet of things: literature review and challenges," *International Journal of Distributed Sensor Networks*, vol. 11, no. 8, p. 431047, 2015.

[2] Z. Feng and Y. Zhu, "A survey on trajectory data mining: techniques and applications," *IEEE Access*, vol. 4, pp. 2056–2067, 2016.

[3] F. Restuccia, N. Ghosh, S. Bhattacharjee, S. K. Das, and T. Melodia, "Quality of information in mobile crowdsensing: Survey and research challenges," *ACM Transactions on Sensor Networks (TOSN)*, vol. 13, no. 4, p. 34, 2017.

[4] P. M. Aoki, R. Honicky, A. Mainwaring, C. Myers, E. Paulos, S. Subramanian, and A. Woodruff, "A vehicle for research: using street sweepers to explore the landscape of environmental community action," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2009, pp. 375–384.

[5] H. Mousa, S. B. Mokhtar, O. Hasan, O. Younes, M. Hadhoud, and L. Brunie, "Trust management and reputation systems in mobile participatory sensing applications: A survey," *Computer Networks*, vol. 90, pp. 49–73, 2015.

[6] X. Kang, L. Liu, and H. Ma, "Data correlation based crowdsensing enhancement for environment monitoring," in *Communications (ICC), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–6.

[7] I. Koukoutsidis, "Estimating spatial averages of environmental parameters based on mobile crowdsensing," *ACM Transactions on Sensor Networks (TOSN)*, vol. 14, no. 1, p. 2, 2018.

[8] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos, "Online outlier detection in sensor data using non-parametric models," in *Proceedings of the 32nd international conference on Very large data bases*. VLDB Endowment, 2006, pp. 187–198.

[9] S. Tasnim, N. Pissinou, and S. Iyengar, "A novel cleaning approach of environmental sensing data streams," in *Consumer Communications & Networking Conference (CCNC), 2017 14th IEEE Annual*. IEEE, 2017, pp. 632–633.

[10] L. Bracciale, M. Bonola, P. Loreti, G. Bianchi, R. Amici, and A. Rabuffi, "CRAWDAD dataset roma/taxi (v. 2014-07-17)," Downloaded from url http://crawdad.org/roma/taxi/20140717, Jul. 2014.

[11] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-air: When urban air quality inference meets big data," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1436–1444.

[12] Y. Zhang, C. Szabo, and Q. Z. Sheng, "Cleaning environmental sensing data streams based on individual sensor reliability," in *International Conference on Web Information Systems Engineering*. Springer, 2014, pp. 405–414.

[13] H.-S. Lim, Y.-S. Moon, and E. Bertino, "Provenance-based trustworthiness assessment in sensor networks," in *Proceedings of the Seventh International Workshop on Data Management for Sensor Networks*. ACM, 2010, pp. 2–7.

[14] K. L. Huang, S. S. Kanhere, and W. Hu, "On the need for a reputation system in mobile phone based sensing," *Ad Hoc Networks*, vol. 12, pp. 130–149, 2014.

[15] D. Peng, F. Wu, and G. Chen, "Pay as how well you do: A quality based incentive mechanism for crowdsensing," in *Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. ACM, 2015, pp. 177–186.

[16] S. Liu, Z. Zheng, F. Wu, S. Tang, and G. Chen, "Context-aware data quality estimation in mobile crowdsensing," in *INFOCOM 2017-IEEE Conference on Computer Communications, IEEE*. IEEE, 2017, pp. 1–9.

[17] Y. Kishino, K. Takeuchi, Y. Shirai, F. Naya, and N. Ueda, "Datafying city: Detecting and accumulating spatio-temporal events by vehicle-mounted sensors," in *Big Data (Big Data), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4098–4104.

[18] J. W. Branch, C. Giannella, B. Szymanski, R. Wolff, and H. Kargupta, "In-network outlier detection in wireless sensor networks," *Knowledge and information systems*, vol. 34, no. 1, pp. 23–54, 2013.

[19] A. Deligiannakis, Y. Kotidis, V. Vassalos, V. Stoumpos, and A. Delis, "Another outlier bites the dust: Computing meaningful aggregates in sensor networks," in *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on*. IEEE, 2009, pp. 988–999.

[20] N. Giatrakos, Y. Kotidis, A. Deligiannakis, V. Vassalos, and Y. Theodoridis, "Taco: tunable approximate computation of outliers in wireless sensor networks," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. ACM, 2010, pp. 279–290.

[21] G. Chen, X.-Y. Liu, L. Kong, J.-L. Lu, Y. Gu, W. Shu, and M.-Y. Wu, "Multiple attributes-based data recovery in wireless sensor networks," in *Global Communications Conference (GLOBECOM), 2013 IEEE*. IEEE, 2013, pp. 103–108.

[22] L. Kong, M. Xia, X.-Y. Liu, M.-Y. Wu, and X. Liu, "Data loss and reconstruction in sensor networks," in *INFOCOM, 2013 Proceedings IEEE*. IEEE, 2013, pp. 1654–1662.

[23] S. Gill, B. Lee, and E. Neto, "Context aware model-based cleaning of data streams," in *Signals and Systems Conference (ISSC), 2015 26th Irish*. IEEE, 2015, pp. 1–6.

[24] S. Krishnan, J. Wang, E. Wu, M. J. Franklin, and K. Goldberg, "Activeclean: interactive data cleaning for statistical modeling," *Proceedings of the VLDB Endowment*, vol. 9, no. 12, pp. 948–959, 2016.

[25] A. Lazar, L. Jin, C. A. Spurlock, K. Wu, and A. Sim, "Data quality challenges with missing values and mixed types in joint sequence analysis," in *Big Data (Big Data), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2620–2627.

[26] H. Liu, A. K. Tk, J. P. Thomas, and X. Hou, "Cleaning framework for bigdata: An interactive approach for data cleaning," in *Big Data Computing Service and Applications (BigDataService), 2016 IEEE Second International Conference on*. IEEE, 2016, pp. 174–181.

[27] S. Tasnim, J. Caldas, N. Pissinou, S. Iyengar, and Z. Ding, "Semantic-aware clustering-based approach of trajectory data stream mining," in *2018 International Conference on Computing, Networking and Communications (ICNC)*. IEEE, 2018, pp. 88–92.

[28] X. L. Dong, B. Saha, and D. Srivastava, "Less is more: Selecting sources wisely for integration," in *Proceedings of the VLDB Endowment*, vol. 6, no. 2. VLDB Endowment, 2012, pp. 37–48.

[29] S. Burdakis and A. Deligiannakis, "Detecting outliers in sensor networks using the geometric approach," in *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*. IEEE, 2012, pp. 1108–1119.

[30] C. Franke and M. Gertz, "Orden: Outlier region detection and exploration in sensor networks," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. ACM, 2009, pp. 1075–1078.

[31] M. Keally, G. Zhou, and G. Xing, "Watchdog: Confident event detection in heterogeneous sensor networks," in *Real-Time and Embedded Technology and Applications Symposium (RTAS), 2010 16th IEEE*. IEEE, 2010, pp. 279–288.

[32] B. Sheng, Q. Li, W. Mao, and W. Jin, "Outlier detection in sensor networks," in *Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing*. ACM, 2007, pp. 219–228.

[33] H. H. Bosman, G. Iacca, A. Tejada, H. J. Wörtche, and A. Liotta, "Spatial anomaly detection in sensor networks using neighborhood information," *Information Fusion*, vol. 33, pp. 41–56, 2017.