A novel data quality assessment framework for vehicular network testbeds

Daxin Tian[1,2,3,4] , Yukai Zhu[1,3,4], Jianshan Zhou[1,3,4], Xuting Duan[1,3,4]*, Yunpeng Wang[1,3,4], Jeungeun Song[5] , Hui Rong[2], Peng Guo[2]

1 Beihang University, Beijing Advanced Innovation Center for Big Data and Brain Computing, XueYuan Road No.37, 100191 Beijing, China
2 China Automotive Technology and Research Center, Automotive Engineering Research Institute, East Xianfeng Road No.68, 300300 Tianjin, China
3 Beihang University, Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems & Safety Control, School of Transportation Science and Engineering, XueYuan Road No.37, 100191 Beijing, China
4 Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, Si Pai Lou. 2, 210096 Nanjing, China
5 School of Computer Science and Technology, Huazhong University of Science and Technology, 430074 Wuhan, China
*Corresponding Author: Xuting Duan

Abstract. Big data technique is considered as a powerful tool to exploit all the potential of the Internet of Things and the smart cities. The development of internet of Vehicles (IoV) and wireless communication technologies have boosted diverse applications related to smart cities and Cyber-Physical Systems, but the data quality of vehicular sensors is an important issue due to the high-speed mobile wireless communication environment and physical sensor noise. This paper presents our experiences for big data analytics based on a vehicular network testbed, in terms of sensors data management, multi-dimension data fusion and data quality assessment for the vehicular sensor data. The proposed data quality assessment framework consist of feature extraction based on multi-sensor data fusion and multi-level wavelet transform, as well as a semi-supervised learning based classification algorithm. The comparison experiment shows that the proposed framework and approaches can extract feasible features and solve the unbalanced label problems, which achieve a better assessment effect.

Keywords: connected vehicles, data quality assessment, data fusion, machine learning, vehicular network

1 Introduction

With the development of wireless communication technologies, Internet of Vehicles (IoV) and data mining techniques, people pay more and more attentions to utilization of the vehicular networks data, which has high worthiness in commercial spheres, city design and traffic engineering. Their flexibility and capabilities are being

TRIDENTCOM 2017, September 28-29, Dalian, People's Republic of China

DOI 10.4108/eai.28-9-2017.2273211

extended, with the infrastructure capacities to provide an Internet access to the cellular networks, and wireless sensors networks through technologies such as Device-to-device (D2D) communication. D2D is an example of an integrated network technology that appears to be a vital component in next generation communication technologies (5G). The D2D technology allows for direct message delivery between terminals that are near each other to lighten the load of node base stations, improve spectral reuse, and enhance system capacity [1]. D2D communication serves as a natural approach to enable reliable and efficient vehicle-to-vehicle (V2V) communication.

Focus on our topic on Internet of vehicle, vehicular network data need to be of high quality to reflect the operation pattern and running status of the vehicle, including vehicle fuel consumptions and travelling route[2]. However, due to abnormalities of sensors and the unreliability of wireless communication environment, some data of vehicular sensors is inaccurate or incomplete. Hence, the quality of vehicular data is difficult to be guaranteed, which has strong relation to the validity and accuracy of data mining results [3][9]. Therefore, quality control and analysis on these data is indispensable, in order to assessment the quality level of the data or to detect the sensor error, filter out the vehicle with inferior data quality.

In this paper, we give a brief introduction of vehicular network big data platform and its data quality issue at first. The database and outliers situation are introduced in section 2. The rest sections are spent reviewing notable examples of automated processing procedures in sensor networks and classification on these preprocessed time-series data. Then we proposed a novel classification framework based on data fusion and machine learning approach to assess the quality of vehicular data. A comparison experiment and detailed procedure of data processing are also given in rest section.

2 Vehicular Network Testbed and Data Quality Assessment

2.1 Vehicular Network Testbed

A brief system framework of integrated vehicular network is shown in Fig.1, which contain cellular networks and V2V communication [4]. The topic and dataset introduced in this paper is from a vehicular network testbed, which contain the vehicle running state data of more than 60,000 vehicles. The vehicles update data on real time which consists of GPS coordinates, speeds, driving direction angle, fuel level value and so on. The real-time big data platform based on the V2I and V2V communication system form the vehicular network testbed. Fig.2 is a screenshot of the real-time big data platform of our vehicular network, the data types of database is also shown.

Vehicular network testbed consists of a large number of sensors and collects multi-dimensional data in high frequency, such as locations, time, velocity and fuel level. However, there are several types of abnormal situation during the process of fuel level data acquisition. The reasons causing outliers in fuel level data can be divided into several types: refueling, fuel spilling or gasoline theft, which will cause the fuel level to decrease rapidly; errors of fuel level sensor or wireless vehicular network transmission [5]; the large shake of vehicle.

Analysis of vehicular network testbed data quality belongs to data quality control. Generally, data quality is defined by correctness, completeness, consistency. Due to the fluctuation of vehicles and precision of fuel level sensors, there are many errors in collected original data by fuel level sensors [6]. During the transmission and storage of fuel level data, stability of wireless vehicular networks and availability of data transformation for storage also cause errors in data [7]. The quality of fuel level data is affected not only by the precision of fuel level sensors but also the quality of transmission networks. There are existing literature discussing the improvement of the precision of fuel level sensors [8] and the quality of transit networks [9].

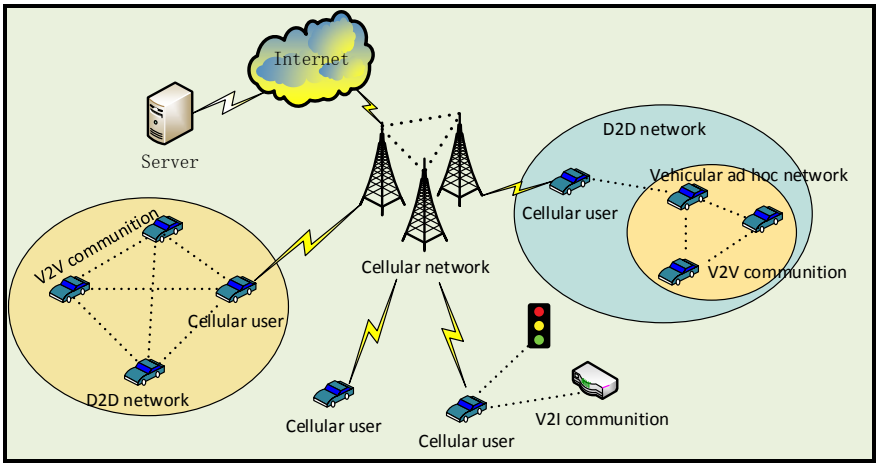

Fig. 1. Integrated vehicular networks

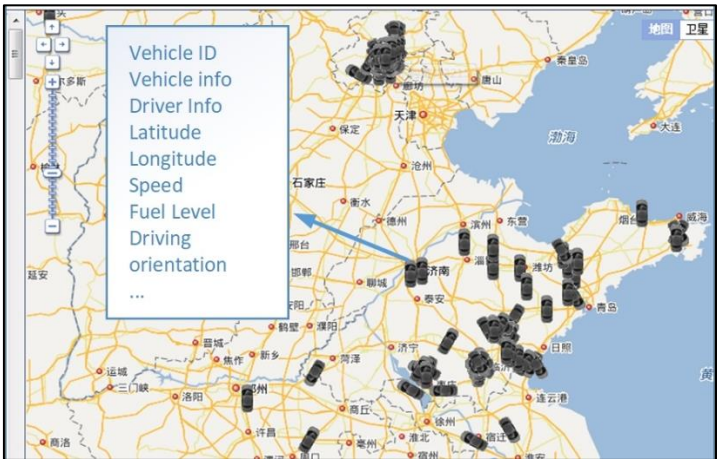

Fig. 2. Vehicular network big data platform

2.2 Classification Based Data Quality Assessments.

Data quality assessments in a further issue based on anomaly detection, which is generally a statistical approach [9]. It could make use of the historical distribution of sensor data. It's important that applying a method that is developed in one area to another is not practicable. Classification-based assessment uses the knowledge of sensed phenomena so as to infer the quality state [10]. As a result, the classification approach is much stronger connected to their application area; this is often achieved by building a suite of rules. For the model used to model the whole system, particularly the functional dependence between states of its sensors. Those systems model the uncertainties of sensor operation but not uncertainties related to the system process given the narrow bounds of its behavior. Environmental process is far more complicated to model because of the variation of contributing factors and location [11]. To assess the data quality of vehicular network, this uncertainty should be typical in the network. So, the proposed framework models not only the process but also operational uncertainties of each sensor.

Since our current focus is on IoV applications scenes, so, there is few work in this special area. The rest of this field is spent reviewing those standing examples of automated quality procedures in the sensor network and anomaly detection in time-series data. When a data sample fails a test, it is labeled as "bad". Such approach fails to use the contextual relationship between the different quality test and the test uncertainty.

3 Proposed Data Quality Assessment Framework and Methodology

3.1 Multi-sensor data fusion.

Multi-sensor data fusion generally provides significant advantages in data mining procession. In addition to the statistical advantage gained by obtaining an improved estimate of a physical phenomenon through redundant observations. the use of multiple types of sensors may increase the accuracy with which a phenomenon can be observed.

Data fusion with vehicular sensor data. In order to assess the data quality of the speed data, we compute the travel speed according to the GPS coordinates, and compare it with the speed which is collected from running cars on real time [10]. The travel speeds could be computed by latitudes φ and longitudes λ, and we compute the relative error ratio of the uploaded speeds v_u with the calculated travel speeds v_t. The relationship could be shown in the Fig.3. We also found the minimum ε of all vehicles in the database is nearly 95 percent which means that the two speeds are approximate extremely, thus we assume that the speed and GPS coordinates are mainly correct as reference values in evaluating the accuracy of the fuel level data. As Fig.3 shows, there is a physical relationship between fuel level and speed value, hence

we use standard deviation of these data and correlation coefficient between these multi-dimension as features for classifier.

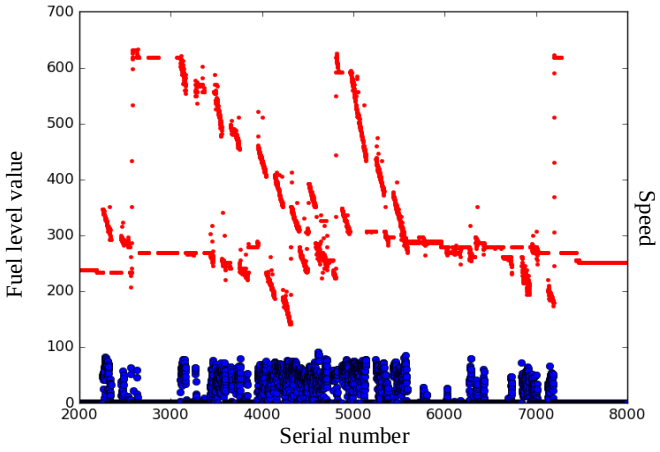

Fig. 3. Distribution of fuel level and speed data.

3.2 Features extraction based on wavelet transform.

In the preprocessing procedure of many types of sensor data, the original series data which may contain variety of noise and anomaly pattern. With the decomposition with Discrete Wavelet Transform (DWT) [13], the set of approximation coefficients A_k retained the main profile of the original data, which is a smooth serious with less noise. On the other hand, the set of detail coefficients D_k represented the noise and slope which contains anomaly pattern in different scaling function. Hence the set D can be analyzed to find out different anomaly pattern. The DWT of a signal x is calculated by passing it through a series of filters. First the samples are passed through a low pass filter with impulse response g resulting in a convolution of the two:

$$y[n] = [n](x * g) = \sum_{k=-\infty}^{\infty} x[k]g[n-k] \tag{1}$$

The signal is decomposed simultaneously using a high-pass filter h. The outputs giving the detail coefficients from the high-pass filter and approximation coefficients from the low-pass. The two filters are known as a pair of mirror filter. However, since half the frequencies of the signal have now been removed, half the samples can be discarded.

$$y_{\mathrm{h}}[n] = [n](x * g) = \sum_{k=-\infty}^{\infty} x[k]\mathrm{h}[2n-k] \tag{2}$$

$$y_{\mathrm{l}}[n] = [n](x * g) = \sum_{k=-\infty}^{\infty} x[k]g[2n-k] \tag{3}$$

Multi-scale wavelet transform. Given a time series X of length n, the DWT consists of $\log_2 n$ stages at most. The first step produces, two vectors of coefficients: approximation coefficients A_1 and detail coefficient D_1. Which are of length $n/2$.

The next step splits the approximation coefficients into two parts using the same scheme, replacing X by A_1, and producing A_2 and D_2, and so on. The wavelet decomposition of the time series at level k has the following structure: $H(X) = A_k, D_1, D_2, ..., D_k$, $|A_k| = 1/2k, |D_j| = 1/2j (j = 1, 2, ..., k), k \le \log_2 n$.

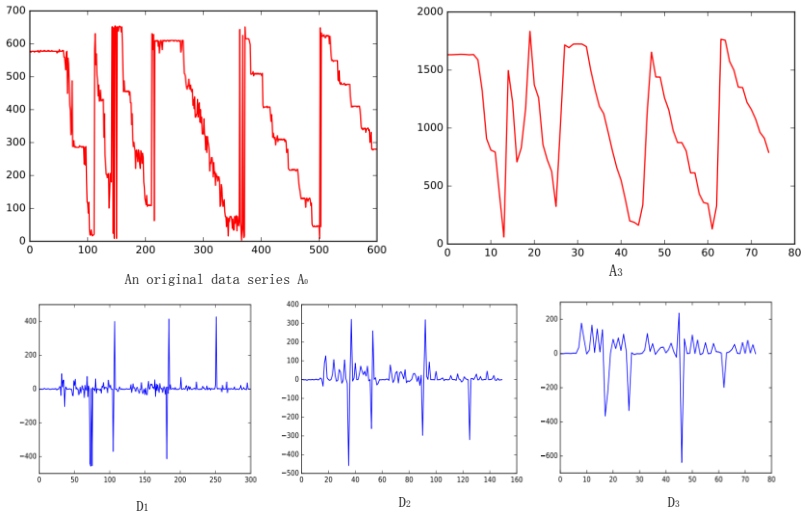

Fig. 4. Multi-level discrete wavelet transform

As the first few wavelet transform coefficient is the strongest after wavelet transform of time series and the highest frequency part of wavelet coefficient include most of the abnormal events and noise, the profile of time series fuel-level data remains basically unchanged when only the first few coefficients are retained. A three level DWT on a series of our case data is shown in Fig.4. Since different coefficients contains information and energy in difference scale, their statistical parameter and energy coefficient are useful and brief features of the data serial.

3.3 Cluster based under-sampling.

Given a representative set of labelled data, a supervised classification approach is an appropriate approach. However, data of bad labels exists in such data sets with low frequency leading to a class imbalance problem. Thus we adopt a cluster based under-sampling method to solve the imbalance problem. For a set S of samples of a particular quality flag (e.g., class) the under–sampling method usually selects instances randomly. A random under–sampling process have the risk of excluding representatives for the original clusters. Hence we use a cluster based random under–sampling pro-

cess that first divides the data in *k* clusters, and randomly select datasets from each cluster.

3.4 Semi-supervised learning classification

Bayesian Network (BN) is a useful and powerful algorithm for uncertainty reasoning. BN includes the Directed Acyclic Graphs for representing the causal relationship between attributes, the conditional probability tables which contain prior probability required for calculating joint probabilistic distribution.

For data set S with attribute set $X = \{X_1, \ldots, X_n\}$ and class label $C = \{c_1, \ldots, c_m\}, t = \{x_1, \ldots, x_n\}$. is a sample in S, NB predicts the class label of t by calculating probability. To overcome the limitation of conditional independent assumption, Functional dependency is an important part of relational database, which represent the constraint relations among different attributes. Existing researches have found some similarities between relational database and probabilistic inference. Then we have:

$$\arg\sum_{\max} P(c \mid x_1, \cdots, x_n) = \arg\sum_{\max} P(c)P(x_1, \cdots, x_n \mid c) = \arg\sum_{\max} P(c)\cdots P(\alpha \mid c)\cdots P(x_n \mid c) \tag{4}$$

In semi-supervised learning, as the joint likelihood of the labeled and unlabeled data is not in closed form, usual solutions to this would be to use Expectation Maximization (EM) [8]. For our case testbeds, a method that updates the model in a continual and an evolutionary manner can be used in the present study. This problem has been researched by some researchers such as Stauffer et al. [11], and Chen et al [9]. The algorithm initials model parameters from limited amount of labeled data and uses these to get probabilistic labels for each unlabeled sample in E-step. M-step gets the parameters using these labels for unlabeled instances. A regulating variable λ is used in the proposed method. This parameter moderates the unlabeled data by reduce the learning rate η and the weight of the unlabeled samples instep M.

$$\begin{aligned} Initial: \theta_L &= \arg_\theta \max \sum_{x \in D_l} \log p_\theta(x, y) \\ Estep: \forall x \in D_u \cup D_l compute p_{\theta_t(y|x)} & \\ Mstep: \theta_{t+1} &= \arg_\theta \max \sum_{x \in D_u \cup D_l} \log p_{\theta_t}(x, y) \\ N_{yi} &= \sum_{x \in D_u} f_i^x P_{\theta_t}(y \mid x) \end{aligned} \tag{5}$$

N_{yi} is denominator in computing the probability of feature f_i being in class y. θt is current estimator of the model parameters. f_i^x gives the count of feature in instance x. Conditional estimates of $P(y \mid x)$ from labeled data improves the accuracy of Bayesian approach.

4 Experimental Results

In the vehicular network data quality assessments QC problem, the quality states of this deployment are adopted from the flagging scheme. There are flags associated with particular processing tasks. The quality flags are designed to indicate the level of uncertainty involved in the sensor reading. Sometimes sensor readings associated with abnormal events may be incorrectly labelled. These labels belong to four different classes of the classification problem. The data that we used in the experiments were labelled into four classes: 1) good data; 2) probably good data; 3) bad data which are potentially correctable, we called it probably bad data; 4)Bad data.

The percentages of instances from different classes (quality flags) in the data sets exists severe class imbalance in the data, hence the under-sampling make a big difference and good effect on the experimental results. We made comparison experiments on three approaches, including straightforward decision, and our proposed framework which are inseparable of whether use class-based under-sampling or random sampling.

In order to test the accuracy of the classifier, we use a 5-cross validation test, which is a way to assess the fitness of a model to a hypothetical validation set when the explicit validation set is not available [13]. At first of the algorithm, we build the training set and testing set, as well as compute the feature vectors. In order to evaluate the classification accuracy, we use F-measure indicators to as the accuracy indicator, which is generally used in machine learning classification problem. F-measure considers both the precision and the recall of the test to compute the score. The F-score can be interpreted as a weighted average of the precision and recall, where an F-score reaches its best value at 1 and worst at 0.

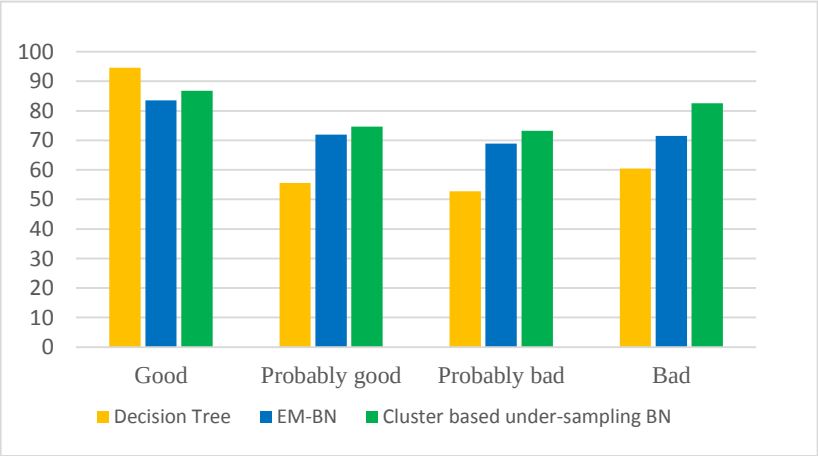

Fig. 5. Comparison of classification accuracy

The classification experimental result is shown in Fig.5, because the unbalanced label problem, the common classification algorithm has low accuracy to classify the bad labelled samples. The proposed framework has a good performance for the average accuracy in each class. For our classification problem, cluster oriented random sampling performs better than the traditional random sampling algorithm. This is due to the fact that the cluster guided sampling approach produces a better approximation of the underlying distribution of data than random under-sampling. The fig.6 shows a labeled data series after separated into many pieces of length, thus a long series of data can be evaluated for each period of time. The blue lines denoted the probably good or bad curves, and the yellow one is series with bad flag.

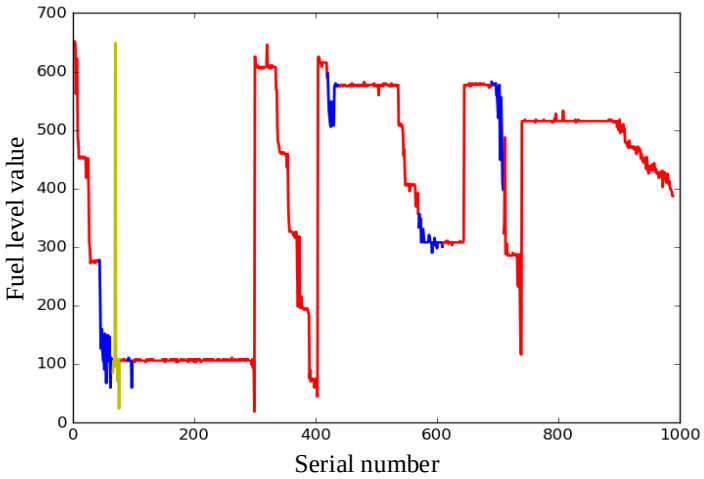

Fig. 6. Processed data series

5 Conclusion

This work presents our experiences for big data analytics based on a vehicular network big data testbed, in terms of sensors data management, multi-dimension data fusion and data quality assessment for the vehicular sensor data. In this paper, we have investigated the problem of multidimensional analysis of vehicular network testbed data. Some statistical indicators are introduced and applied in data quality evaluation, and a novel classification based framework is proposed to efficiently assess the data quality and screen out the abnormal vehicles in database. Moreover, . Our experiments on large real datasets show the feasibility and practical utility of proposed framework and approaches.

We will also develop the further work of data analysis to support more complex applications, such as the atypical event detection and knowledge discovery based on

vehicular network. There are still many challenges in designing an appropriate information dissemination method and building a strong and reliable vehicular network platform as well as the big data testbeds. With the developing data mining and intelligent computing techniques, the potential of the huge amount of vehicular network data would be exploit, which will contribute construction of future Cyber-Physical Systems and green smart city.

Acknowledgments. This research was supported by the National Key Research and Development Program of China (2016YFB0100902).

References

1. Jara AJ, Varakliotis S, Skarmeta AF, Kirstein P. Extending the internet of things to the future internet through IPV6
2. support. Mobile Information Systems 2014; 10(1):3–17.D. Tian, et al., "Optimal epidemic broadcasting for vehicular ad hoc networks," Int. J. Commun. Systems, Vol. 27, no. 9, Sep 2014, pp. 1220-1242.
3. Hou, X., Li, Y., Chen, M., Wu, D., Jin, D., Chen, S.: Vehicular fog computing: A viewpoint of vehicles as the infrastructures. IEEE Transactions on Vehicular Technology 65(6), 3860–3873 (2016)
4. K. Hwang, M. Chen, Big Data Analytics for Cloud/IoT and Cognitive Computing, Wiley, U.K., ISBN: 9781119247029, 2017.
5. Chen, S., Hu, J., Shi, Y., Zhao, L.: Lte-v: A td-lte-based v2x solution for future vehicular network. IEEE Internet of Things Journal 3(6), 997–1005 (2016)
6. Chen, S., Zhao, J.: The requirements, challenges, and technologies for 5g of terrestrial mobile telecommunication. IEEE Communications Magazine 52(5), 36–43 (2014)
7. Ahn, K., Rakha, H., Trani, A., Van Aerde, M.: Estimating vehicle fuel consumption and emissions based on instantaneous speed and acceleration levels. Journal of Transportation Engineering 128(2), 182–190 (2002)
8. Christopher, M.B.: Pattern recognition and machine learning. Company New York Ny 16(4), 049901 (2006)
9. Chen, M., Hao, Y., Kai, H., Wang, L., Wang, L.: Disease prediction by machine learning over big data from healthcare communities. IEEE Access 5(99), 8869–8879 (2017)
10. D. Smith, G. Timms, and P. de Souza: A quality control frameworkfor marine sensing using statistical, causal inference, in Proc. IEEE OCEANS, Kona, HI, USA, Sep. 2011, pp. 1–7.
11. Benvenuto, F., Marani, A.: Neural networks for environmental problems: Data quality control and air pollution nowcasting. Global Nest the International Journal (2000)
12. Elnahrawy, E., Nath, B.: Poster abstract: online data cleaning in wireless sensor networks. In: International Conference on Embedded Networked Sensor Systems. pp. 294–295 (2003)
13. Nicewander, J.L.R.W.A.: Thirteen ways to look at the correlation coefficient. The American Statistician 42(1), 59–66 (2012)
14. Tang, L.A., Yu, X., Kim, S., Han, J., Peng, W.C., Sun, Y., Gonzalez, H., Seith, S.: Multidimensional analysis of atypical events in cyber-physical data. In: 2012 IEEE 28th International Conference on Data Engineering. pp. 1025–1036 (2012)