ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

An Optimal Linear Predictive Model for Missing Data Estimation in Wireless Sensor Network

Doreswamy, Yogesh K M

Department of Computer Science, Mangalore University, Mangalagngotri, Karnataka, India

Department of Computer Science, Mangalore University, Mangalagngotri, Karnataka, India

ABSTRACT: In dealing with wireless sensor networks, the missing data from sensor is foreseeable due to the limited constraints of nodes and causes many problems in various applications. To solve such problems, the missing data should be estimated as accurately as possible. In this paper, an optimal linear predictive model is proposed for estimating the missing sensor data. This model is designed on the temporal data sets of sensor nodes closure to a target sensor node generating missing data. The proximity of target sensor node and its spatial neighboring nodes is determined by K-Nearest Neighbor Algorithm. The proposed liner predictive model is experimented on Berkley Intel Lab datasets and results show that the proposed algorithm can estimate the missing data accurately. Various error measures are evaluated to validate the performance of the liner predicting model. The best one is suggested for building liner predictive model for the estimation of missing data in the wireless sensor network.

KEYWORDS: Wireless Sensor Network, Missing Data Estimation, KNN Algorithm, Linear Predictive Model, Error Measures.

I. INTRODUCTION

The aggressive developments of sensor technology, wireless communication technology and computer technology have revolutionized the wireless sensor network (WSN) technology. Investigation on WSN becomes a hotspot research area and is extensively used in various application fields [2][8]**.** However, there are limitations in a sensor node that include energy limitation in energy, communication range, processing and storing capability. A sensor node in wireless sensor network can't work when it runs out of its limitation, or easily impacted by the circumstance, or it cannot save many sensor data in its memory. In all the uncertain events, the data that could be generated from each sensor node can be considered as missing data or some time that may lead to outlier data[13][17]**.**

Actually, the missing of sensor data is inevitable due to the intrinsic properties of WSNs. Firstly; the communication ability of sensor nodes is restricted. A number of sensor nodes may be detached from the WSNs for a short or long time due to the influences of surrounding location like mountains and obstacles, which results that the sensor data of these nodes may be lost. Furthermore, more effects comes from the natural environment like thunder, rain, and lightning will influence the sensor nodes' communication quality and affects on the links of communication between sensor nodes connected and disconnected frequently. This will also result in the missing data of sensor in the time of the data transmission.

Secondly, the power of sensor nodes is limited. When a sensor node's power is low down, it sometimes works under an unstable state. This not only causes the unstable communication, which may resulting in losing of the sensor data , but also makes the sensor data sampled be often useless abnormal data (e.g. the temperature of a room is 300^{o}C). The abnormal data is looked as the missing data since it can never be used. The practical problems of this kind can be seen from the data sets obtained by the wireless sensor network deployed in Intel-Berkeley joint laboratory [11] and in the red wood by university of California, Berkeley [16]**.**

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

The rest of the paper is organized as follows: section II describes the related works on missing data estimation in Wireless Sensor Network. Section III describes the design aspects of Linear Regression Model, section IV describes the Pseudo code of the proposed method, and experimental results are discussed in section V. Conclusions are given in section VI.

II. RELATED WORK

Query processing is one of the crucial tasks in WSNs. It deals with continuous queries and approximate queries in order to reduce time complexities [1]**.** Processing continuous queries is dynamic task as the data streams out from a node and streams in to another node in WSN. Scheduling continues queries to collect the sensing data that satisfy the queries depend on the network topology and the characteristics of other systems such as spatial relationships and as well as temporal information of nodes [18]**.** Optimizing continuous queries converges mainly on how to utilize the temporal-spatial relationship of sensing data through statistical/ mathematical models that answer the queries approximately and minimize communication cost. Missing values in sensing data decline the performance of query processing by increasing the processing time and make the mathematical models to answer approximate queries wrongly towards decision making. Therefore missing data estimation in WSNs is critical problem in real environment.

Missing data in real time applications is inevitable and leads to many problems during data analysis. Therefore, missing data estimation a critical problem in every filed where decision making is based on data analysis outcome. Research on missing data estimation has been carried out in many fields, such as Bioinformatics [14]**.** Physical and statistical methodologies were developed for Data estimation methods in sensor networks [9]**.** The idea of k-nearest neighbor algorithm is proposed for missing data estimation on the temporal and spatial correlation of sensor data [10]**.** The K-Nearest Neighbor classifier was proposed for material classification in Materials Informatics [4]**,** pattern classification in Artificial Intelligence [12]. Data imputation in compound descriptors data set in Chemo informatics [3] and Data Mining [1]. A Study on Error Approximation Methods for Predictive Regression Model is done [5]**.** It adopts the linear regression model to describe the spatial association of sensor data among distinct sensor nodes, and utilizes the data information of multiple neighbor nodes to predict the missing data jointly instead of independently, so that a stable and reliable estimation performance can be achieved [13].

III. PROPOSED ALGORITHM

In many real applications, the environment variables such as temperature, humidity and, voltage changes continuously and these are monitored by WSN. The changes in these variables are generally due to variations in the natural environment. When some data of a sensor node is missed, a naive method for estimating the missing data is based on the non-missing temporal correlation of sensor data.

However, this method works well only when the sensor data changes smoothly and the missing data comes in to view in short time duration. In the other cases, this method may cause large estimation errors. This becomes as a result of the sensor data in WSNs changes sharply and irregularly often in fact, especially the data sensed in the natural environment because there are too many uncertain factors, such as environment noise, will affect the variety of the sensor data. So, only depending on the temporal correlation of sensor data to estimate the missing data is not enough in many cases. Therefore, spatial relationships of the sensor nodes and their temporal correlated data are considered to estimate the missing data of a sensor node. K-Nearest Neighbour Algorithm (K >= 2) is proposed for considering sensor nodes, whose location is closure to sensor node. The temporal information of the K-NN sensors are considered for building liner regression model and estimating missing data of a sensor node [10] [17]**.**

A. K-Nearest Neighbour (K-NN) Algorithm

K-Nearest Neighbour (KNN) A**lgorithm** is part of machine learning[12] that has been used in many applications in the field of data mining[16][6]**,** statistical pattern recognition and many others. The closure of two sensors is computed by Euclidian distance function [15]**.**

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

Sensor nodes in a region of interest or area are identified by their location, which is represented special coordinate, (x_i, y_j), the distance between any two sensor nodes , $S_1(x_i, y_j)$ and $S_2(x_i, y_j)$ is computed using Euclidian distance function, $d(S_1, S_2) = \sqrt{(x_{2,i} - x_{1,i})^2 + (y_{2,j} - y_{1,j})^2}$, (1)

For a sensor with missing data is identified as target sensor node say S2, whose nearest neighbor sensors could be, K ≥3, say S3, S4 and S5, otherwise there would be much deviated sensors that maximizes model error.

B. Correlation Coefficient(r)

Correlation gives an indication of the strength and direction of a linear relationship between two variables-dependent(Y) and independent(X) variables. There are a lot of different coefficients are used for different situations. The best known is the Pearson product-moment correlation coefficient (also called Pearson correlation coefficient or the sample correlation coefficient), which is computed by dividing the covariance of the two variables by the product of their standard deviations. If there are *n* number observation of these variables, $X = \{x_1, x_2, x_3 \ldots\ldots x_n\}$ and $Y = \{y_1, y_2, y_3 \ldots\ldots x_n\}$, then the Pearson product-moment correlation coefficient can be used to find the correlation between two variables.

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}} \quad (2)$$

Where $\bar{x}$, and $\bar{y}$ are the means of n observations of indent dependent variable(X) and dependent variable (Y) respectively.

The correlation is +1 in the case of a perfect increasing linear relationship, and -1 in case of a decreasing linear relationship, and the values in between indicates the degree of linear relationship between X and Y. A correlation coefficient of 0 means the there is no linear relationship between the variables. The square of the Pearson correlation coefficient (r^2), known as the Coefficient of Determination, describes how much of the variance between the two variables is described by the linear fit.

C. DESIGN CONSIDERATIONS

Linear regression model is designed with two steps. In the first step, a set of sensors nodes generated missing data is identified. In the second step, for a given sensor node having missing data, K-Nearest neighbour sensor nodes are selected by K-Nearest Neighbour Algorithm. The non-missing data sets of the nearest sensor nodes are considered for building linear regression mode. The data generated by the sensor nodes are generally non-linear in nature. It requires seeing the linear relationship of the sensor data sets of sensors for building model. Therefore, correlation analysis deployed to find the degree of relationship. A set of sensor data, whose degree of relationship, r ≥ 0.9, is considered for building model.

Linear Regression Model : The data collected by the sensor S_i at time t is represented as data sets, D_{n*m}={s1(a1,a2,a3,a4...am), t1), (s2(a1,a2,a3,a4...am),, t2), (s3(a1,a2,a3,a4...am), t3), …, sn-1(a1,a2,a3,a4...am), tn-1)}, containing n number of instances with unique time stamp, ti, and each instance containing *m* number of attributed values(nature of data obtained say in general Temperature, Humidity, Light and Voltage, Pressure, Force.. etc.). The degree of relationship, r ≥ 0.9 of two attributed variable vectors, (x, y), for instance linear relationship exists between Temperature and Humidity, is determined for selecting model variables. Let $x = \{x_1, x_2, x_3 \ldots\ldots x_n\}$ be an independent variable and $y = \{y_1, y_2, y_3 \ldots\ldots x_n\}$ as dependent variable. The mathematical model that has the functional linear relationship between these variables and fits to a straight line is defined

$y = f(x) = a + bx$ Where a is a slope and b is an intercept. (3)

This line equation is the linear mathematical model for the data, where a and b are model parameters. Therefore, it is required to find the optimal values of the model parameters a and b that will best fit the experimental data [[6] [7].

Scalar parameters, (a, b), of model are computed for the entire data set and used as a measure for comparing model quality in procedures of searching to discover the best regression model for the data. For a known structure of

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

the linear regression model y = a+ bx, optimal model parameters values have to be computed based on the given data sets $x = \{x_1, x_2, x_3 \ldots\ldots x_n\}$, and $y = \{y_1, y_2, y_3 \ldots\ldots x_n\}$ and the defined performance criterion. Finding values of parameters is the static minimization problem and these parameters can be defined as below:

$$a = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} x^2{}_i - \left(\sum_{i=1}^{N} x\right)^2}, \qquad b = \frac{N\sum_{i=1}^{N} x_i{}^2 \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i y_i}{N\sum_{i=1}^{N} x^2{}_i - \left(\sum_{i=1}^{N} x\right)^2} \quad (4)$$

Regression Errors: A model for data sets is an approximation of the reality; computation of a predicted outcome will come with certain error. Let us consider the i[th] data point (x_i,y_i) from the data set D(X,Y), with values x_i and y_i of the independent and dependent variable, respectively; the predicted variable value computed by the model $y_i^{est} = a + bx_i$ lies on the regression line. The estimated value y_i^{est} deviated from the true data value y_i from the data point (x_i, y_i). This difference, e = real-value − predicted-value, i.e, is the **regression error** also called the **residual** or **modelling error**. It is expressed as follow:

$$e_i = (y_i - y_i^{est}) = y - (a + bx_i) \quad (5)$$

D. Performance Criterion:

The performance of regression model is measured by the performance criteria. There are various measures such as Root Mean Square Error (RMSE), Normalized Root Mean Square (NRMS), Mean Absolute Error (MAE), Sum of Square Error (SSE)[7][15].

a). **Mean Absolute Error (MAE)** : $E_{mae} = \frac{1}{N}\sum_{i=1}^{N} \left| \hat{y}_i - y_i \right|$ (6)

b). **Mean Squared Error (MSE):** $E_{mse} = \frac{1}{N}\sum_{i=1}^{N} \left(\hat{y}_i - y_i \right)^2$ (7)

c). **Sum of Square Error (SSE):** $E_{SSE} = \sum_{i=1}^{N} e(i)^2 = \sum_{i=1}^{N} (y_i - \hat{y})^2$, where $\hat{y} = f(x_i) = a + bx_i$ (8)

d). **Relative Absolute Error (RSE):** $E_{rae} = \frac{\sum_{i=1}^{N} \left| \hat{y}_i - y_i \right|}{\sum_{i=1}^{N} \left| y_i - \bar{y} \right|}$ (9)

e). Relative Squared Error (RSE): $E_{rae} = \frac{\sum_{i=1}^{N} \left| \hat{y}_i - y_i \right|}{\sum_{i=1}^{N} \left| y_i - \bar{y} \right|}$ (10)

a. **Root Mean Square Error (RMSE):** $E_{rmse} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} \left[\hat{y}_i - y_i \right]^2}$ (11)

Where y_i is observed values and $\hat{y}_i$ is modeled values at time/place *i*.

f). **Normalized Root Mean Square Error (**NRMSE**):** $E_{nrms} = \sqrt{\frac{\sum_{i=1}^{N} \left[\hat{y}_i - y_i \right]^2}{N \sum_{i=1}^{N} \left[\hat{y}_i - \bar{y} \right]^2}}$ (12)

IV. PSEUDO CODE

The general steps for generating model for estimating missing data of a sensor node is as follows:

Inputs: Sensor Node Locations(x, y) and their data instances

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

Output: An optimal Linear Regressing Model, $y = f(x) = a + bx$, with parameter values

Methods:

1. Determine the sensor node generating missing data, S_{nmd}.
2. Determine the parameter K, number of nearest neighbour nodes beforehand, around the missing data node.
3. Distance Computation: Actual distance from the sensor node generating missing data say, $S_{msd,}$ to each other nodes using Euclidian distance algorithm. $d(S_{nmd}, S_i)$.
4. Sort all the nodes in non-decreasing of their distance proximity and determine the K-nearest neighbour nodes.
5. K-NN Algorithm: Select K-NN sensor nodes, around node, $N_{msd,}$ for building predictive Model.
6. Determine Model parameters, dependent variable(Y) and independent variable(X) were determined among the sensor data reading types, light, humidity, temperature and voltage. Correlation among these determined and perfect correlated attributed data sets are considered for model building.
7. Generate linear predictive model for estimating missing data of sensor node, $N_{msd,}$ Humidity(Y) = a *Temperature (X) + b. Compute optimal values of a and b for different sets of varying data instances.
8. Evaluate the model with different Error Measure
9. Select the error Measure that minimizes the error.

V. EXPERIMENTAL RESULTS

Design and Development of any model on any data set is induced for validating in order to ensure the performance for the implementation of the model in real environment. The proposed linear regression model for the estimation of missing data of Wireless Sensor Network was implemented on the data sets of WSNs **(S. Madden, 2014),** which is a trace of readings from 54 sensor nodes deployed in the Intel Research Berkeley lab. These sensor nodes collected light, humidity, temperature and voltage readings once every 31 seconds. Data generated from these 54 sensor nodes contain 2.3 million data instances and stored in the database. Functional modules were built for Missing Data Sensor Node Selection, distance computation model, K_NN model, Model building and error estimation models. An optimal predictive Model, Humidity(Y) = - 0.35 *Temperature (X) + 47.42 was built with data instances, **709677** of maximum 17 sensor nodes nearer to node 5, which is having the largest number of missing values.

Different error measures are compared based on other normalized error measure values. Figure 1 plots number of instances of each group K along X-axis and their normalized error values by different sets instances of data, along Y-axis. The maximum error value is initially started from 1.0 and exponentially decreasing towards 0. The minimum error value by MSE is converged to 0.00. Therefore, the MSE method is the best Error approximation method that can be considered for measuring the predictive Linear Regression Model for WSN data instances.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

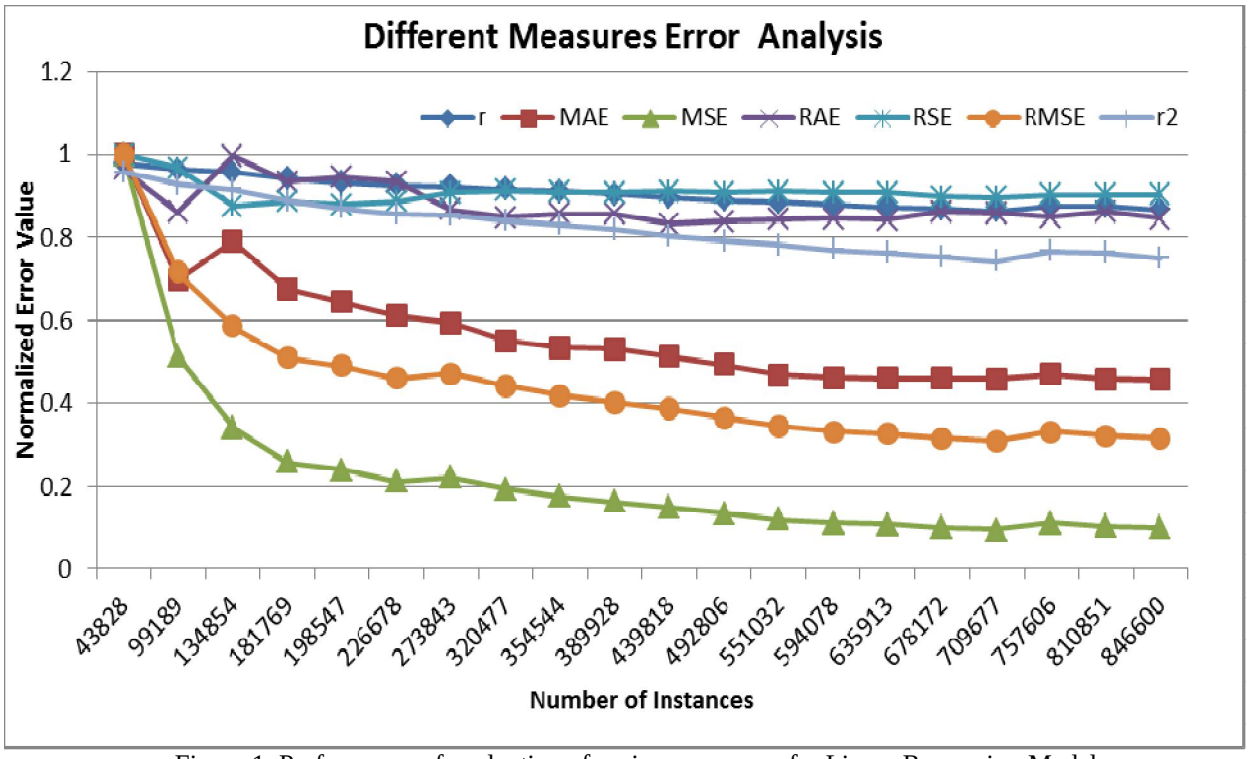
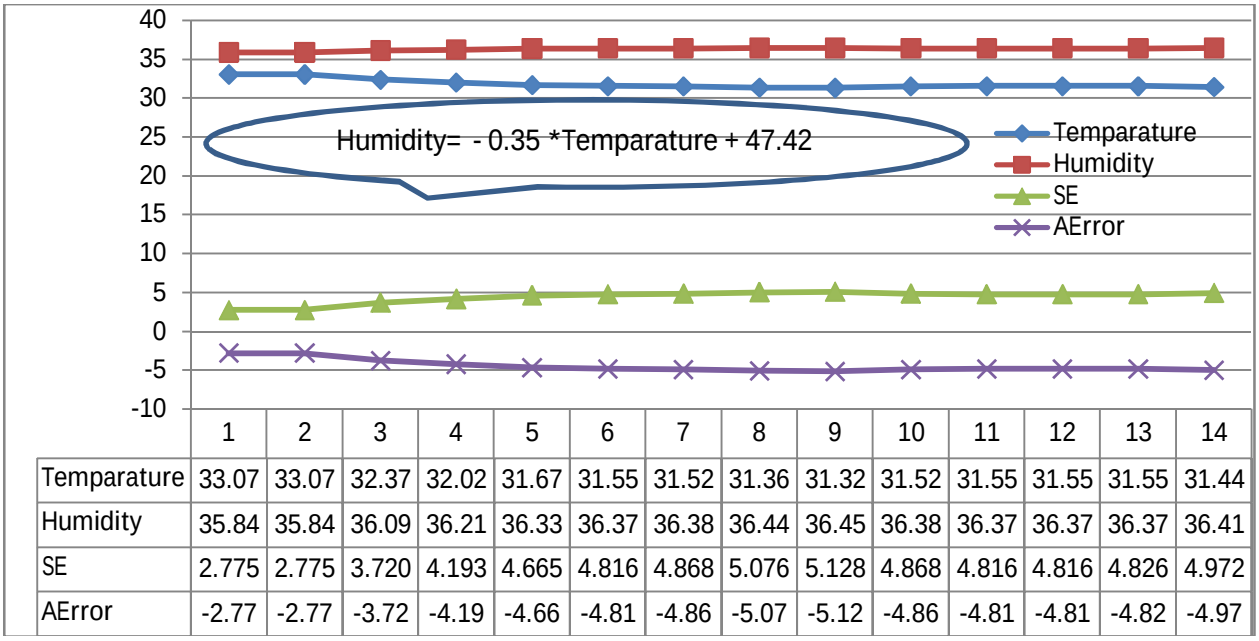


Figure 1: Performance of evaluation of various measures for Linear Regression Model



	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Temparature	33.07	33.07	32.37	32.02	31.67	31.55	31.52	31.36	31.32	31.52	31.55	31.55	31.55	31.44
Humidity	35.84	35.84	36.09	36.21	36.33	36.37	36.38	36.44	36.45	36.38	36.37	36.37	36.37	36.41
SE	2.775	2.775	3.720	4.193	4.665	4.816	4.868	5.076	5.128	4.868	4.816	4.816	4.826	4.972
AError	-2.77	-2.77	-3.72	-4.19	-4.66	-4.81	-4.86	-5.07	-5.12	-4.86	-4.81	-4.81	-4.82	-4.97

Figure 2: Profiles of Predicted Values of Humidity from Temperature values along with minimized error

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

From the figure 2, it shows that for any value of temperature, model predicts the actual humidity value with the actual error (AError). The computed errors (Square error) and the actual errors (AError) are similar and equal. The proposed model behaves perfectly on any input data sets. There should always be some negative correlation between the temperature and humidity.

VI. CONCLUSION AND FUTURE WORK

In this paper, an optimal predictive liner regression model for estimating the missing data of WSN is proposed and experimented. Here, different estimation methods, Correlation Coefficient(r), Mean Absolute Error(MAE), Mean Square Error(MSE),Relative Absolute Error(RAE),Relative Square Error(RSE), Root Mean Squared Error(RMSE) and Coefficient of Determination(r2) were evaluated on the predictive Linear Regression Model , Humidity(Y) = - 0.35 *Temperature (X) + 47.42, designed for K-NN group, K=17, containing the sensor nodes with 709677 data instances. The MSE method is found be the best among the other methods. Hence, it is proposed as the best error approximation method for validating the predictive linear regression model built for data sets. Further this model can be extended to multilinear prediction model and it can be compared with other prediction models.

REFERENCES

1. AzharMahmood, Ke Shi, Shaheen Khatoon, and Mi Xiao (2013) ,Data Mining Techniques for Wireless Sensor Networks: A Survey, International Journal of Distributed Sensor Networks, REviw artcle,,pp. 1-24, http://dx.doi.org/10.1155/2013/406316.
2. Cullar D, Est rin D, Strvastava M. (2004): Overview of sensor networks. IEEE Computer, Vol. 37, issues 8, pp.41-49.
3. Doreswamy and Chanabasayya .M. Vastrad (2013). A Robust Missing Value Imputation Method Mifoimpute for Incomplete Molecular Descriptor Data and Comparative Analysis with other Missing Value Imputation Methods, International Journal on Computational Sciences & Applications (IJCSA) Vol.3, No. 4, pp. 63-74.
4. Doreswamy, Hemanth.K.S (2011). "Hybrid Data Mining Technique for Knowledge Discovery from Engineering Materials Data sets" International Journal of Database Management Systems (IJDMS) Vol.3, No.1, pp.166-177.
5. Doreswamy, Srinivas, Narasegowd, (2014), A Study on Error Approximation Methods for Predictive Regression Model, International Journal of Computational Intelligence System, France, Atlanta Press, France. Under review by peer team for publication.
6. Jiawei Han, Micheline Kamber and Jian Pei,(2012): Data Mining Concepts and Techniques, 3rd, edition,
7. Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski, Lukasz A. Kurgan,(2010). Data Mining a Knowledge Discovery Approach, Springer Publications.
8. Li Jianzhong, Gao Hong (2008). Advance of wireless sensor networks, Journal of Computer Research andDevelopment. Vol. 45, issue 1, pp.1-15.
9. Li Y, Ai C, Deshmukh W P, et al.(2008). Data estimation in sensor networks using physical and statistical methodologies[C], Proc of the 28th IEEE Int Conf on Distributed Computing Systems. Washington: IEEE Computer Society, pp. 538-545.
10. Liqiang Pan, Jianzhong Li (2010) ,K-Nearest Neighbor Based Missing Data Estimation Algorithm in Wireless Sensor Networks , Wireless Sensor Network,, Vol.2, 115-122, doi:10.4236/wsn.2010.22016 y 2010 (http://www.SciRP.org/journal/wsn/).
11. Madden S. Intel Berkeley research Lab data. [2006-08-08]. http://berkeley. Intel-research. Net/lab data accessed in January 2014.
12. Mohammad Abu Alsheikh, Shaowe Lin, Dusit Niyato and Hwee-Pink Tan, (2014), Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications, accepted for publication IEEE Journal is 2004 Issues.
13. Pan Liqiang, Li Jianzhong, Luo Jizhou(2010). A Temporal and Spatial Correlation Based Missing Values Imputation Algorithm in Wireless Sensor Networks. Chinese Journal of Computers, Vol.33, issues 1, pp.1-11.
14. Sehgal, Gondal, Dooley, and Coppel(2008), "Ameliorative missing value imputation for robust biological knowledge inference," Journal of Biomedical Informatics, Vol. 41, No. 4, pp. 499–514, 2008.
15. Sudha, Doreswamy, (2009). Prototype Engineering Selection System (PEMSS)-Similarity Measuring Approach, Chapter II: A Novel Distance Metric for Engineering Materials Selection, MSc IV Semester Project, Department of Computer Science, pp. 12-20.
16. Sweta Kumari and Varsha Singh,(2011): A Comprehensives Survey Paper on Sensor Data Mining Based on Sensor, International Journal of Technology, Vol. 1, Issue 1, pp. 37-41.
17. Tolle G. Sonoma redwoods data. [2006-08-08].http://www.cs.berkeley.edu/～get/ sonoma, 2014.
18. Xiaozhen YAN, Hong XIE, Tong Wang (2011). A Multiple Linear Regression Data Predicting Method Using Correlation Analysis for Wireless Sensor Networks, Journal of Computational Information Systems, Vol 7, issues 11, pp. 105-4112.
19. Yang, H. B. Lim, M. T. Ozsu, and K. L. Tan. (2007). "In- network execution of monitoring queries in sensor networks," In SIGMOD Conference, Beijing, China, pp. 521–532.