DATA REPOSITORY FOR SENSOR NETWORK: A DATA MINING APPROACH

Doreswamy and Srinivas Narasegouda

Department of Computer Science, Mangalore University, Mangalore, India

ABSTRACT

The development of sensor data repositories will aid the researchers to create benchmark dataset. These benchmark dataset will provide a platform for all the researchers to access the data, test and compare the accuracy of their algorithms. However, the storage and management of sensor data itself is a challenging task due to various reasons such as noisy, redundant, missing, and faulty data. Therefore it is very important to create a data repository which contains the precise and accurate data and also storage and management of data is effective. Hence, in this paper we are proposing to use the combination of quantitative association rules and decision tree for classification of faulty data and normal data. Usage of multiple linear regression models for the estimation of missing data. A symbolic table approach for storage and management of sensor data. And development of a graphical user interface for visualization of sensor data.

KEYWORDS

Data Mining, Association Rules, Decision Tree, Sensor Network, Faulty data, Missing Data, Multiple Linear Regression, Sensor Data Repository.

1. INTRODUCTION

Research and development in technology has resulted in producing various electronic devices which are aiding humans in many ways. Out of all devices, sensor is one such device which plays a major role in modern day humans' life. The applications of sensors include many fields such as environmental monitoring, habitat monitoring, wild life monitoring, fire alarm system, disaster management, surveillance systems and many more. In many applications set of sensors are deployed in an area to set up a sensor network. These sensor networks consisting of small devices capable of sensing, acquiring, processing and transmitting the data to the base station where user can study them and make significant decisions. Sensor networks generate a massive amount of data in the form of stream in a quick real time. Storage and management of such a huge data itself is a challenging task. Apart from this, sensor devices are prone to errors due to various reasons such as low battery power, environmental effects, hardware malfunctioning etc. This kind of errors results in producing faulty data. Using such faulty data for analysis or for decision making may result in getting ineffective, untrustworthy results. Hence, in order to get the accurate result, faulty and inaccuracy in the data must be removed before it is being considered for analysis. And once the faulty data are removed, precise, accurate data needed to be stored in an efficient way so that retrieval and processing of data can be made easier and effective. Many techniques have been developed in the past by various researchers in order to remove faulty data which are mentioned in the next section.

The rest of the paper is structured as follows. Section 2 gives the survey of literature; Section 3 describes the proposed model. Section 4 explains the experimental results and conclusion is given in Section 5.

DOI : 10.5121/ijdms.2014.6401

2. Literature Survey

Data in the real world may consist of noisy, faulty, redundant, and missing data due to various reasons. Hence, before storing the data into the data repository, it becomes mandatory to make sure that data to be stored in the repository is precise and accurate. Detection of faulty, noisy data has been a challenging task for many researchers. In [1], statistical techniques were combined with cross-validation technique to identify the online sensor faults. By exploiting the spatiotemporal relations and using those in Bayesian algorithm faults were detected [2]. In [3], external entity such as external manager was used for fault detection. But the communication between the sensor node and the external manager results in declining the energy of sensor nodes. Statistical measure such as median was used in [4] for confirming whether it is a faulty data or normal data by comparing the median value with the nearest neighbours. In order to reduce the communication cost by using the least number of neighbours and to develop a fault tolerant mechanism, Bayesian and Neyman-Pearson method was used in [5]. Fuzzy data fusion mechanism [6] and fuzzy classifier mechanism [7] were also developed for fault detection in sensor network. By studying the data centric view, statistical and environmental features, different types of faults can be identified [8].

Once the faulty data is removed, data need to be stored in the data repository. In the past few years many researchers have proposed many techniques for storage and management of sensor data. A German Federal Environmental Agency proposed a XML based Environmental Markup Language for representing the environmental data [9]. Sensors are playing an important role in autonomous driving. A research was conducted by collecting the data using NIST High Mobility Multi-purpose Wheeled Vehicle and data was stored in a relational database [10]. The concept of internet blog was also used in order to develop SensorBase.org data repository. And similar to SensorBase.org, CRAWAD is another web based sensor data repository [11]. An object oriented approach was also used in the past to develop sensor data repository [12-14].

3. Proposed Model

The main objective of the proposed model is to develop a data repository for sensor network which contains precise and accurate data. In this matter the proposed model is divided into five modules such as (i) symbolic table for storage and management of sensor data. (ii) Generation of association rules for quantitative data. (iii) Classification of expected and unexpected data using quantitative association rules. (iv) Estimation of missing data using multiple linear regression. And finally, (v) graphical user interface to provide visualization model for the data repository.

3.1. Symbolic Table for Storage and Management of Sensor Data

Sensor network generate the data in the form of streams. And over the period of time, distribution of data changes in the stream [15]. This change in the distribution of data indicates the spatiotemporal relations among the sensors. And, it is important to note that consideration of historical data for analysis may affect the spatiotemporal relations. Hence, it becomes inevitable to store the data in such a way that spatiotemporal relations are not affected. For this, we are proposing a model to store the data in a symbolic table rather than in traditional databases.

The detailed explanation about how the traditional data can be converted into symbolic data is explained in [16]. In our previous work [14] we have used an object oriented approach to develop a sensor data repository in which data was stored in a symbolic table. In order to create a category, we used three variables namely, *sensor id, maximum number of days,* and *date*. And the equation used for generating a key for storage and management of data is as follows

$$Key_{ij} = (S_i-1)(max\_day)+D_j \quad (1)$$

Where, S_i is the set of sensors i.e., $S_i= \{S_1, S_2, S_3, . . . ,S_n\}$. Sensor id represents the physical positioning of the sensor i.e., space dimension. *max_day* is the maximum number of days for which sensor network was set up, and D_j is the set of all distinct dates i.e., $D_j= \{D_1, D_2, D_3, . . . ,D_n\}$. Date attribute D_j also represents the time dimension. In this approach all the sensor readings generated by a sensor on in a day were stored under one key. But from literature we know that sensor data changes over a period of time. Hence, we modified the equation (1) to deal with change of distribution of data over a period of time. For this, we divided the day into equal time intervals of six i.e., data generated in every 4 hours is considered as one interval. The time span between the time on which data is generated and the time at which the network was set up is indicated by $T_{interval}$ and the new equations are as follows.

$$T_{interval} = (\text{Current\_time} - \text{Start\_time}) / T_{max\_ip} \qquad (2)$$

$$Key_{ij} = (S_i\text{-}1)\ (\text{max\_day})\ (T_{max\_ip}) + T_{interval} \qquad (3)$$

Whenever a data is generated, data is stored into the symbolic table using the equation (3).

3.2. Association Rules for Quantitative Data

The main objective of data mining techniques is to extract the knowledge from the abundance of raw data. Association rule plays a vital role in the extraction of knowledge from the data. Association rules are mainly used for categorical data. However, data is not restricted to categorical only. Data can also be quantitative in nature. Based on the statistical theory, [17] developed a quantitative association rule technique. The association rules generated by this technique are in the form *population-subset* → *extraordinary-behaviour*. For example if we have two features temperature and humidity the generated association rules will be in the form of *(temperature1, temperature2)*→ *average (humidity).* This technique has its own drawbacks. First, on the left hand side it contains only one quantitative attribute. Secondly, applying this technique to sensor network data may not be feasible because the distribution of sensor data changes over a period of time. Hence, in order to deal with we added two attribute namely time and space attribute on the left hand side of the association rule using the equation (3).

Our algorithm for generating the association rule is similar to [17] except that in our approach the left hand side of the association rule contains three attributes namely space, time, and temperature compared to only one attribute in [17]. And the algorithm is as follows.

Input: An array D of transactions attributes space, time, temperature, humidity, and a value mindif. D is sorted according to the attribute key which contains space, time, and temperature.

Output: Association rules for key to humidity

```
Window (D, key, humidity, mindif)
        Window-above (D, key, humidity, mindif)
        Window-below (D, key, humidity, mindif)
```

The procedure for *Window-above ( ), Window-below ( ),* and definition for the variable *mindif* is same as in [17].

3.3. Classification of Data using Decision Tree

The structure of the decision tree is similar to flowchart consisting series of *If Then Else* conditions. In decision tree, each non-leaf node indicates a test on an attribute. The outcome of test is represented by a branch. And each terminal node or leaf node holds the class label.

In our approach we are proposing to use decision tree to classify the data into two class namely normal data and faulty data. For classification of data we used the association rules. The flow of data in decision tree is represented in figure 1.

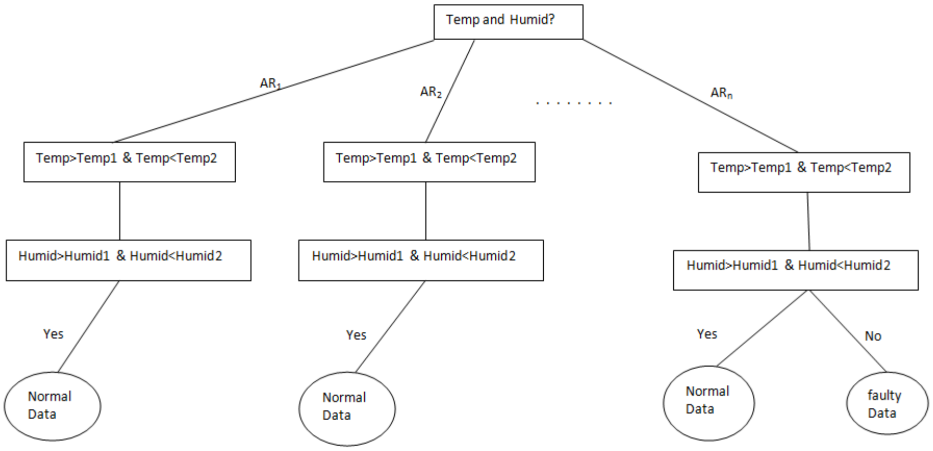


Figure 1 Decision tree for the classification of faulty data.

3.4. Estimation of Missing Data using Multiple Linear Regressions

Missing data in a large database is inevitable. The presence of missing data declines the performance of the data mining techniques. Hence, it becomes mandatory to estimate the missing data before applying any data mining technique. There are numerous methods available for handling and estimation of missing data and multiple linear regression is one such powerful technique. The multiple linear regression equation consists of two variables namely predictor variables X_1, X_2, X_3, . . . ,X_p and dependent variable Y. The regression model as explained in [16], is defined by the following equation (4).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + e \qquad (4)$$

The regression design matrix X is the *n(p+1)* matrix

$$X = \begin{pmatrix} 1 & X_{11} & X_{12} & X_{13} & \cdot & \cdot & X_{1p} \\ 1 & X_{21} & X_{22} & X_{23} & . & . & X_{2p} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & X_{n1} & X_{n2} & X_{n3} & . & . & X_{np} \end{pmatrix}$$

The regression coefficient vector β is the *(p+1)* vector

$$\beta' = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p),$$

And the vector error e is given by

$$e' = (e_1, e_2, e_3, \ldots, e_n)$$

Where error terms satisfy E $(e_i)=0$ and Var $(e_i)=\sigma^2$ and Cov $(e_i,e_{i'}) = 0$, $i \neq i'$.

If X is a non singular then the least squares estimators of the parameter β are given by

$$\hat{\beta} = \left(X'X\right)^{-1} X'Y \qquad (5)$$

3.5. Visualization of Data using Graphical User Interface

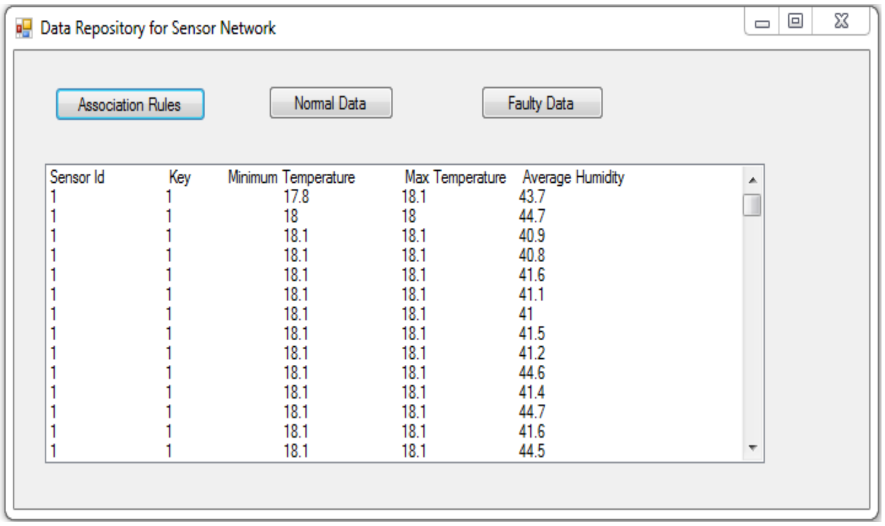
Sensor data are complex in nature. Understanding and analysis of sensor data requires high domain knowledge. Because of this requirement, end users sometimes face a great difficulty in decision making. Hence, we propose to develop a graphical user interface which will aid the user in understanding and analysis of the sensor data.

4. EXPERIMENTAL RESULTS

We have implemented the proposed work using C#. For the experiment purpose we have used the publicly available Intel Berkeley Research lab dataset [18]. This dataset consists of 2.3 million observations with attributes date and time, epoch, sensor id, temperature, humidity, light, voltage. The dataset also contains the information about the physical positioning of each sensor. We conducted our experiment in five phases namely creation of symbolic table, sample data for generating association rules, classification of data using association rules as a condition in decision tree, estimation of missing data, and visualization of data.

In order to deal with the change in the distribution of data over the time, we created the symbolic table using the equation (3) for storage and management of data. We have used dictionary class in C# for implementing the symbolic table.

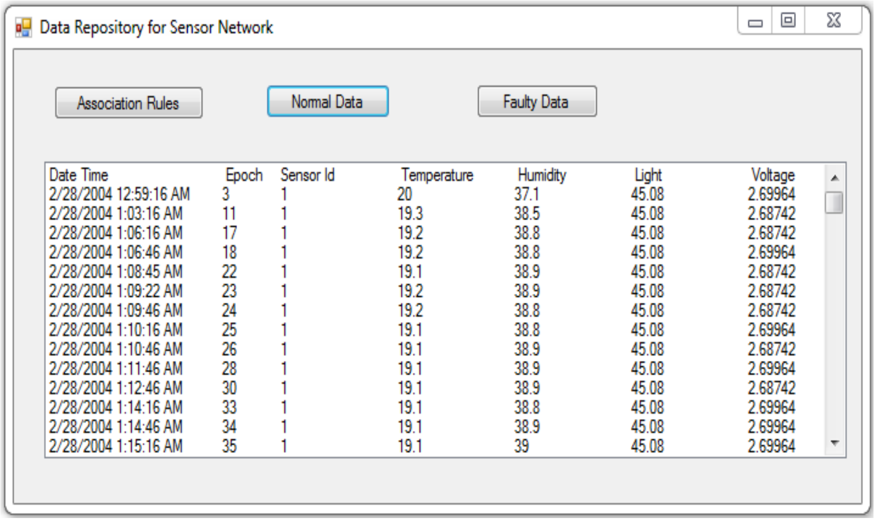
We considered the sample data, collected in the first eight days to generate the association rules for quantitative data. In the sample data, we ignored any missing data present because the missing data may decline the accuracy of the generated association rules. Figure 2 shows the association rules generated for the sensor data.



Sensor Id	Key	Minimum Temperature	Max Temperature	Average Humidity
1	1	17.8	18.1	43.7
1	1	18	18	44.7
1	1	18.1	18.1	40.9
1	1	18.1	18.1	40.8
1	1	18.1	18.1	41.6
1	1	18.1	18.1	41.1
1	1	18.1	18.1	41
1	1	18.1	18.1	41.5
1	1	18.1	18.1	41.2
1	1	18.1	18.1	44.6
1	1	18.1	18.1	41.4
1	1	18.1	18.1	44.7
1	1	18.1	18.1	41.6
1	1	18.1	18.1	44.5

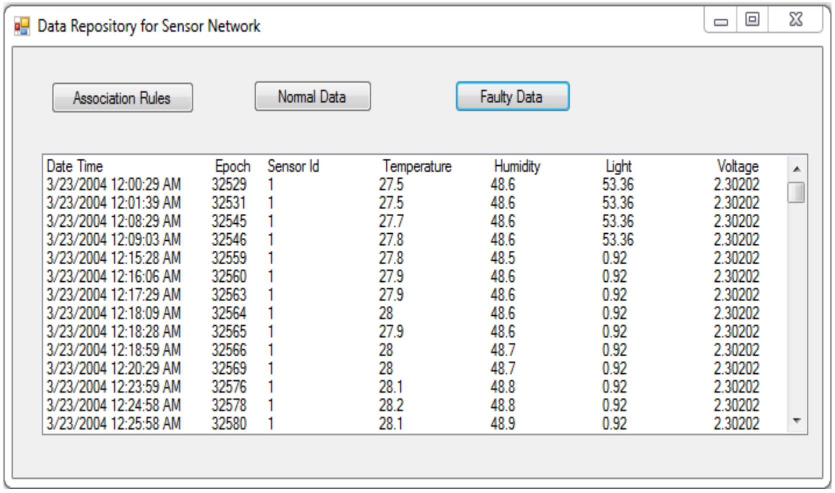
Figure 2 Association Rules for Quantitative Data

Once the association rules are generated, we used the decision tree structure to classify the normal or expected data and faulty data. In decision tree, association rules are placed in the internal nodes as the condition for classification of data. Any data which comply with decision tree conditions are labelled as normal data and those data which do not comply with the decision tree conditions are labelled as faulty data. Figure 2 and figure 3 show the classified normal and faulty data respectively. And table 1 show the results obtained after the classification of data into two classes such as normal data and faulty data.



Date Time	Epoch	Sensor Id	Temperature	Humidity	Light	Voltage
2/28/2004 12:59:16 AM	3	1	20	37.1	45.08	2.69964
2/28/2004 1:03:16 AM	11	1	19.3	38.5	45.08	2.68742
2/28/2004 1:06:16 AM	17	1	19.2	38.8	45.08	2.68742
2/28/2004 1:06:46 AM	18	1	19.2	38.8	45.08	2.69964
2/28/2004 1:08:45 AM	22	1	19.1	38.9	45.08	2.68742
2/28/2004 1:09:22 AM	23	1	19.2	38.9	45.08	2.68742
2/28/2004 1:09:46 AM	24	1	19.2	38.8	45.08	2.68742
2/28/2004 1:10:16 AM	25	1	19.1	38.8	45.08	2.69964
2/28/2004 1:10:46 AM	26	1	19.1	38.9	45.08	2.68742
2/28/2004 1:11:46 AM	28	1	19.1	38.9	45.08	2.69964
2/28/2004 1:12:46 AM	30	1	19.1	38.9	45.08	2.68742
2/28/2004 1:14:16 AM	33	1	19.1	38.8	45.08	2.69964
2/28/2004 1:14:46 AM	34	1	19.1	38.9	45.08	2.69964
2/28/2004 1:15:16 AM	35	1	19.1	39	45.08	2.69964

Figure 3 Classified as Normal Data using Decision Tree



Date Time	Epoch	Sensor Id	Temperature	Humidity	Light	Voltage
3/23/2004 12:00:29 AM	32529	1	27.5	48.6	53.36	2.30202
3/23/2004 12:01:39 AM	32531	1	27.5	48.6	53.36	2.30202
3/23/2004 12:08:29 AM	32545	1	27.7	48.6	53.36	2.30202
3/23/2004 12:09:03 AM	32546	1	27.8	48.6	53.36	2.30202
3/23/2004 12:15:28 AM	32559	1	27.8	48.5	0.92	2.30202
3/23/2004 12:16:06 AM	32560	1	27.9	48.6	0.92	2.30202
3/23/2004 12:17:29 AM	32563	1	27.9	48.6	0.92	2.30202
3/23/2004 12:18:09 AM	32564	1	28	48.6	0.92	2.30202
3/23/2004 12:18:28 AM	32565	1	27.9	48.6	0.92	2.30202
3/23/2004 12:18:59 AM	32566	1	28	48.7	0.92	2.30202
3/23/2004 12:20:29 AM	32569	1	28	48.7	0.92	2.30202
3/23/2004 12:23:59 AM	32576	1	28.1	48.8	0.92	2.30202
3/23/2004 12:24:58 AM	32578	1	28.2	48.8	0.92	2.30202
3/23/2004 12:25:58 AM	32580	1	28.1	48.9	0.92	2.30202

Figure 4 Classified as Faulty Data using Decision Tree

Table 1. Details of data classification

Total number of dataset	2303285	
Total number of missing data	93206	4.0466%
Total number of normal data	1921870	83.4403%
Total number of faulty data	288209	12.5129%

After the classification of data as normal or expected data and faulty data, normal or expected data is used to create the regression model using the equations (4) and (5). Figure 5 shows the graph plotted between the actual value and the estimated value of the attribute temperature. Our experimental result on a sample data shows that, multiple linear regression for the estimation of missing data has a mean absolute percentage error of 7.08% which results in an accuracy of 92.92%. Once the missing data are estimated, estimated data is inserted into the symbolic table using the equation (3). And finally, the visualization of the workflow and results are represented through a graphical user interface.

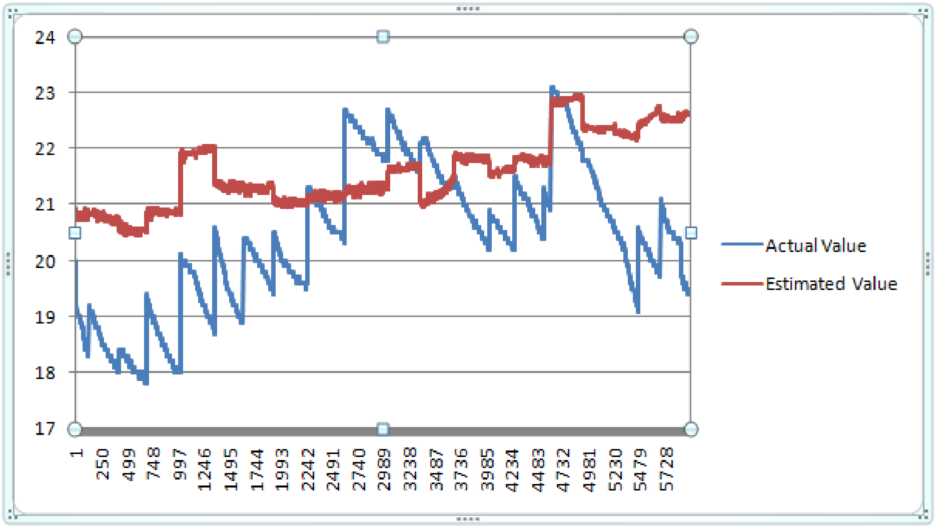


Figure 5 Actual value vs Estimated value

5. CONCLUSIONS

The development of a sensor data repository has a great importance for the researchers. It provides the precise and accurate data for the researchers to test the accuracy of their algorithms on a benchmark dataset. Data in the real world is not always precise and accurate. It may contain noisy, redundant, missing and faulty data. Testing on such inaccurate data may result in inaccurate solutions. Hence, it is very important to eliminate faulty data and estimate the missing data before storing into the data repository. In this paper, we are proposing to use the combination of quantitative association rules with decision tree structure for detection of faulty data. Estimation of missing data using multiple linear regression. A symbolic table was created for the storage and management of sensor data. Finally, graphical user interface was used for visualizing the sensor data. The experimental result shows that faulty data have been successfully detected and removed.

REFERENCES

[1] F. Koushanfar, M. Potkonjak, and Sangiovanni-Vincentelli, (2003) “On-line fault detection of sensor measurements”, *in Sensors*, IEEE, Vol. 2, pp974-979.

[2] B. Krishnamachari and S. Iyengar, (2004) “Distributed bayesian algorithms for fault-tolerant event region detection in wireless sensor networks”, *Computers, IEEE Transactions on*. IEEE, Vol. 53, pp241-250.

[3] L. Ruiz, I. Siqueira, H. Wong, J. Nogueira, A. Louereiro, (2004) “Fault management in event-driven wireless sensor networks”, in *Proceedings of the 7th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems*. ACM, pp149-156.

[4] M. Ding, D. Chen, K. Xing, and X. Cheng, (2005) “Localized fault-tolerant event boundary detection in sensor networks”, in *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies*. IEEE, Vol. 2, pp902-913.

[5] X. Luo, M. Dong, and Y. Huang, (2006) “On distributed fault-tolerant detection in wireless sensor networks”, *Computers, IEEE Transactions on*. IEEE, Vol. 55, pp58-70.

[6] J. Shell, S. Coupland, and E. Goodyer, (2010) “Fuzzy data fusion for fault detection in wireless sensor networks”, in *Computational Intelligence (UKCI), 2010 UK Workshop on*. IEEE, pp1-6.

[7] A. Lemos, W. Caminhas, and F. Gomide, (2011) “Adaptive fault detection and diagnosis using an evolving fuzzy classier”. *Information Sciences*. Vol. 220, pp64-85.

[8] K. Ni, N. Ramanathan, M. Chehade, L. Balzano, S. Nair, S. Zahedi, E. Kohler, G. Pottie, M. Hansen, and M. Srivastava, (2009) “Sensor network data fault types”, *ACM Transactions on Sensor Networks*. Vol. 5, pp25.

[9] H. Arndt, T. Bandholtz, O. Gunther, M. Ruther, and T. Schutz (2000) “Eml-the environmental markup language”, in *Proceedings of the Workshop Symposium on Integration in Environmental Information Systems*.

[10] M. Shneier, T. Chang, T. Hong, G. Cheok, H. Scott, S. Legowik, and A. Lytle (2003) “Repository of sensor data for autonomous driving research”, in *Proceedings of SPIE*. Vol. 5083, pp390-395.

[11] CRAWDAD. http://crawdad.cs.dartmouth.edu/

[12] A. Bauer, T. Emter, H. Vagts, and J. Beyerer (2009) “Object oriented world model for surveillance systems”, in *Future Security: 4th Security Research Conference*. Fraunhofer Verlag, pp339-345.

[13] Y. Fischer and A. Bauer (2010) “Object-oriented sensor data fusion for wide maritime surveillance”, in *Waterside Security Conference (WSS), 2010 International*. IEEE, pp1-6.

[14] Doreswamy, and S. Narasegouda (2014) “Symbolic Data Analysis for the Development of Object Oriented Data Model for Sensor Data Repository”, in *Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA)*. Springer, pp435-442.

[15] N. Jaing and L. Gruenwald, (2006) “Research issues in data stream association rule mining”, *ACM Sigmod Records*. Vol.35, pp14-19.

[16] L. Billard and E. Diday (2006) “Symbolic data analysis: conceptual statistics and data mining”, Wiley.

[17] Y. Aumann and Y. Lindell (2003) “A statistical theory for quantitative association rules”, *Journal of Intelligent Information Systems*. Springer, pp255-283.

[18] Intel Berkeley Research lab dataset, http://db.csail.mit.edu/labdata/labdata.html

AUTHORS

Doreswamy received B.Sc and M.Sc Degree in Computer Science from University of Mysore in 1993 and 1995 respectively. Received Ph.D degree in Computer Science from Mangalore University in the year 2007. After completion of his Post-Graduation Degree, he subsequently joined and served as Lecturer in Computer Science at St. Joseph’s College, Bangalore from 1996-1999.Then he has elevated to the position Reader in Computer Science at Mangalore University in year 2003. He was the Chairman of the Department of Post-Graduate Studies and research in computer science from 2003-2005 and from 2008-2009 and served at varies capacities in Mangalore University. At present he is the Chairman of Board of Studies and Professor in Computer Science of Mangalore University. His areas of Research interests include Data Mining and Knowledge Discovery, Artificial Intelligence and Expert Systems, Bioinformatics, Molecular modelling and simulation, Computational Intelligence, Nanotechnology, Image Processing and Pattern recognition. He has been granted a Major Research project entitled “Scientific Knowledge Discovery Systems (SKDS) for Advanced Engineering Materials Design Applications” from the funding agency University Grant Commission, New Delhi, India. He has published about 30 peer reviewed Papers at national/International Journal and Conferences. He received SHIKSHA RATTAN PURASKAR for his outstanding achievements in the year 2009 and RASTRIYA VIDYA SARASWATHI AWARD for outstanding achievement in chosen field of activity in the year 2010.

Srinivas Narasegouda received his BSc degree from Gulbarga University in 2006 and MSc degree from Karnatak University in 2008. At present, he is pursuing his PhD under the guidance of Professor Doreswamy in the department of computer science Mangalore University, Mangalore. His areas of interest are Data Mining and Knowledge Discovery, Swarm Intelligence, and Symbolic Data Analysis.