Context Aware Similarity Measure Selection: Mining of Wearable Implantable Body Sensor Network Data with Logical Reasoning

Y Indu , V.Uma Maheswari
Assistant Professor, Vardhaman College of Engineering, Hyderabad
Indu12.cse@gmail.com
Assistant Professor, Vardhaman College of Engineering, Hyderabad
umasridhar11@gmail.com

Abstract:

Wireless sensor networks monitor the environment with various types of sensors. Environment in its broader terms can be the geographic environment or it can be our human body. One such type of network is Wearable and Implantable Body Sensor Network (WIBSN). This paper focuses on processing of data generated from WIBSN. WIBSN includes a network of sensors that generate different type of values. This paper treats each sensor as a dimension in the whole dataset. In this case, data may have both continuous and discrete values. Hence; proposed work can be applicable for both of those data values. By identifying nature of the sensor data model, underlying similarity or dissimilarity measure is selected. A novel Crisp clustering technique is used to simulate the proposed work.

Keywords: Wireless Sensor Network, Wearable body sensor, Crisp Cluster, Elastic Clustering

I. INTRODUCTION

Wireless sensor network is a collection of homogenous or heterogeneous sensors. The type of sensor used depends on various parameters such as signal strength needed, frequency of data generation, deployment area etc. These sensors can be controlled and monitored remotely thus enabling of integration of real-world objects with computing systems for making smarter and efficient decision making. There are various industries that require sensors such as Manufacturing, Processing, Retail, and Automotive, Health care, Transport agencies, entertainment and many more. Among these application areas, health industry is having major requirement of sensors and it dependent applications. According to different surveys [1] over the next 10-15 years, focus shifts from providing healthcare to providing quality healthcare at affordable costs, primarily due to increased population among developing countries which is a challenge in the current context.Challenges faced by the stakeholders across the business value chain in Healthcare domain (from Pharma company to Insurance provider) include the need for faster product innovation cycles or reduced Time to market (Pharma companies), increasing health care operational costs, patient behavior and sentiment data (Healthcare providers), and increasing claims cost (for insurance companies). Most of these challenges can be addressed using advanced analytics using integrated data [2].Existing fields of healthcare such as synthetic biotechnology can be used as communication systems for smart health care environments [3]. Such type of environments consists of wearable and implantable body sensor network systems. The objective of this technology is to continuously monitor people during their daily activities by integrating sensing and consumer electronics.

Another application of Wearable body sensor networks includes provision for pre-medication. This can be achieved through providing clinical services, monitoring, and communication in emergency situation using clinical data access. According to proverb "Prevention is better than Cure", successive generations can be prevented from various health problems if the problem is identified and treated in current generation in advance. Processing of such type of data is possible through data mining techniques.

II. RELATED WORK

Embedded Body Sensor Networks: Significant development has happened in wireless sensor networks and biology leading to embedded body sensors that can communicate without affecting the nature of human body i.e. any side effects. The sensors which are implantable in the body are called as biosensors. They can be used to record various health parameters such as pulse, blood pressure, electrocardiograms, body temperature etc. Implantation of sensors within the body can continuously monitor bodily functions and providing aninstant health service based on the data generated from those sensors.

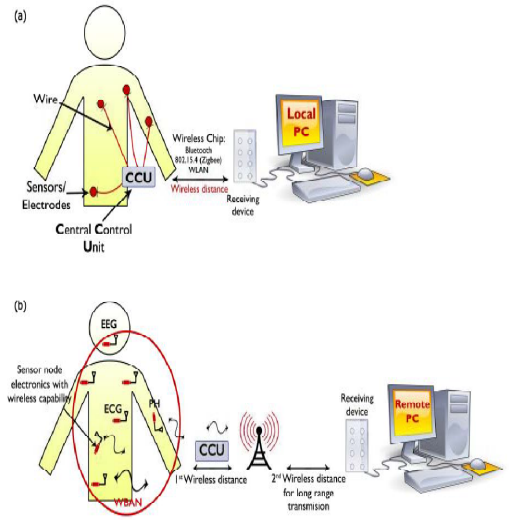

Fig:1 Body Monitoring System (a) Traditional Model (b) Modern Model

According to the above figure (1a) traditional body monitoring systemshave sensors/electrodes connected physically and limited to local monitoring system. But modern systems (figure 1b) use advanced infrastructure to monitor remotely by communicating with the body sensors through powerful wireless channels providing real time and any time patient health data gathering.

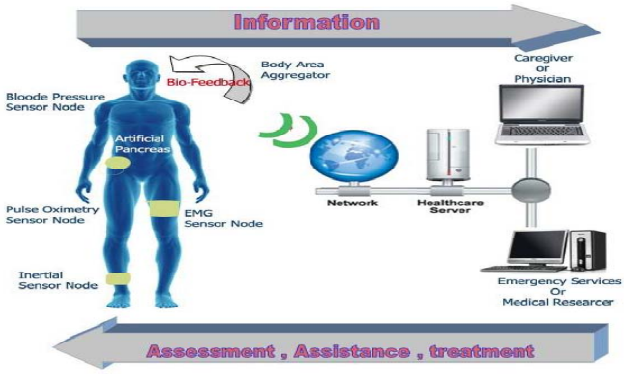

Fig:2 Architecture of Wearable and Implantable Wireless Sensor networks

The above architecture shows how the patient and health care information systems are connected to each other. The data is stored in the health care server. Any health analytics has to use data from that health care server. Most of the existing data mining researches are only applicable and process such data by considering part of the data i.e. either numerical or categorical but not on both at same time and didn't collectively find and analyze the results.

III. PROPOSED SYSTEM

The main objective of the proposed system is to apply the various distance or similarity measures on the data collected from sensors of WIBSN network and clustering the data based on collective results taken from various distance or similarity measuresfor different type of data. We assume that each sensor name is one dimension and belongs to one domain in WIBSN dataset. We are proposing one novel algorithm named as "Context Aware Adaptive Elastic Clustering Algorithm". The algorithm take mixed attribute dataset as input and output clusters based on collective analysis of mixed attributes. Here mixed attributes means both continuous and discrete attributes. Proposed algorithm uses following distance measures.

*Euclidian distance***:** This is very popular distance measure suitable for numerical continuous attributes. The formula to find distance between two instances X and Y is

$$d(X,Y)=\sqrt{\frac{\sum_{i=1}^{n}(X_i-Y_i)^2}{n}}$$

*Jaccard distance***:** This is used to find the distance between two categorical attributes. The formula to find distance two instances X and Y is

$$J_d(X, Y)=\frac{|X\cup Y|-|X\cap Y|}{|X\cup Y|}$$

*Cosine Similarity***:** This measure is used find the similarity between two document vectors i.e this measure is meant for neither numerical nor categorical. If the data or attribute is having any text information then this measure is used. In proposed is case this is used to compare two patients' disease history documents or doctor prescriptions. The formula for two document vectors X and Y is and distance of two vectors based on above formula is

$$\text{sim}(X,Y)=\frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}}$$

$$d(X,Y)= 1/\text{sim}(X,Y)$$

Proposed Clustering algorithm as framed as follows:

Proposed algorithm belongs to Crisp clustering technique, which means any instance must be placed in one cluster only. For this algorithm input is the dataset generated from wearable body sensors. Each sensor serves as an attribute in the dataset. This dataset is a time series dataset also because each sensor reads the value periodically and it is treated as an event. But no two sensors may have same time interval for reading. Hence to find the distance, dynamic time wrapping is used to efficiently handle phase shift and map the sensor readings. Also, numerical attribute values are not in same scale. In that case normalization is required. Normalization is calculated based on the following formula for a given attribute

Procedure Normalize(X)

1. Find the min and max values of X
2. For each value 'V_i' in X the equivalent normalized value is $\frac{(Vi-min)}{(max-min)}$

Algorithm: Context Aware Adaptive Elastic Clustering Algorithm

Input: WIBSN Dataset with Attribute Markers, three distance thresholds

Output: Clusters

1. Dataset Transformation using Normalization
2. Consider first instance as center of the first cluster
3. Move to next instance and compare that instance with existing clusters centers with given distance threshold as follows
   3.1 Partition the given dataset into three parts based on attribute types that is part-1 for numerical attributes,part-2 for categorical and part-3 for text based attributes
   3.2 Find the distance of instances concurrently for all three parts.
4. Now compare these distances of current instance and cluster center with given three distance thresholds.
5. If all the three distances are <= thresholds then
6. 7.1 Place the current instance in the current cluster If current instance is not fit in any cluster then 8.1 Form new cluster with current instance as cluster center
7. Repeat steps 3-8 for all the instances in the dataset
8. Find the quality of each cluster using Davies-Bouldin Index
9. Finally Output the Clusters

The main advantage of above algorithm is, there is no need to specify number of clusters in advance. It automatically computed based on the tuning parameter distance threshold. Whenever you adjust that threshold number clusters are generated. But why this algorithm took three distance thresholds? The answer is simple. It uses three different similarity measures for three types of attributes. Each distance measure result is not compatible single threshold. So, based on the resultant distance range one can tune the thresholds. Other advantage of this algorithm is Adaptive and Elastic nature i.e. the same algorithm can be applicable not only for static dataset but also for data streams. No need to change single line of code in the main algorithm. But this is left as my future work.

According to the above algorithm first record or instance is treated as cluster center such that first cluster is created with single instance and it is center for that cluster. From next record onwards each instance is compared with existing cluster centers created till now. Here algorithm has to find the distance between current instance and cluster center. Internally the distance function finds distance separately for numerical, categorical and text type of data based on distance measures previously described in this paper.

Now three types of distances are generated instance function and those distances are compared with the supplied distance thresholds respectively. If all of the three conditions satisfied then current instance is placed in the cluster with minimum distance. This point is not mentioned above. But the thing is current instance may fit into more than one cluster based on threshold. In that case it has to find minimum distance cluster. It increases quality of the cluster. Finally cluster quality is evaluated through popular Davies-Bouldinindex [4]. This

measure is used to find inter and intra cluster similarity of cluster members i.e. how well the cluster members are scattered within the cluster and how well two clusters are separated are computed using this index.

IV. SIMULATION RESULT

We used diabetic dataset for simulation purpose with 5 sensor readings.

Sample Screens:

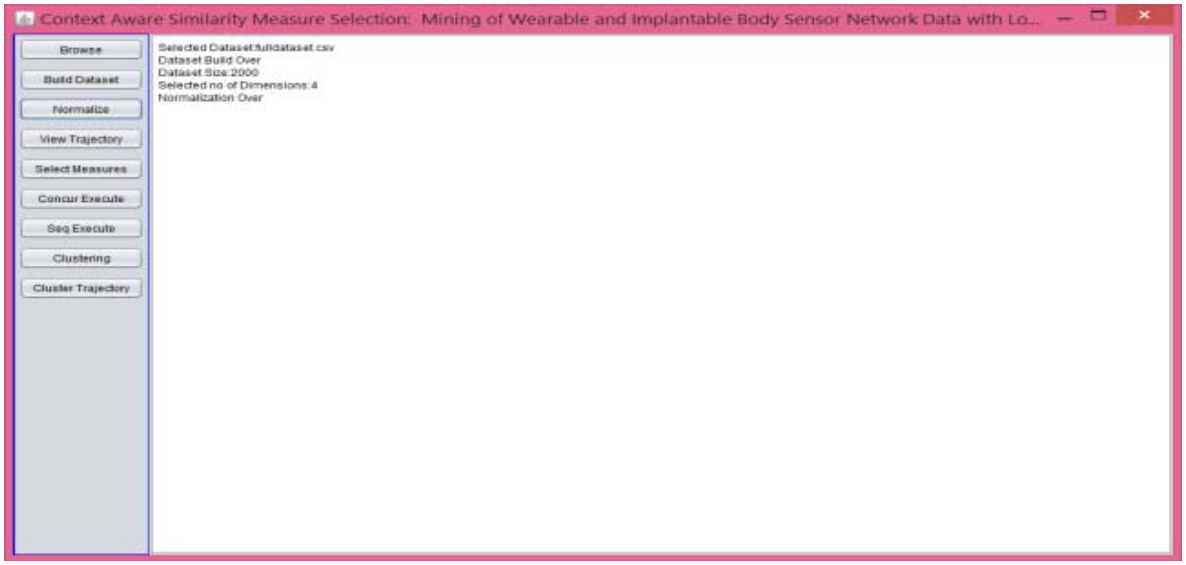

Fig: 3 Context Aware Similarity Measure Selection:

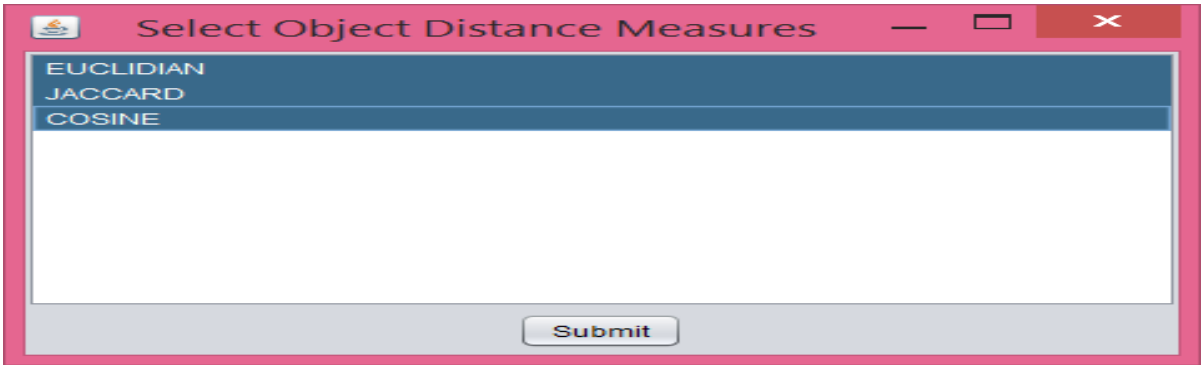

Fig: 4 Object Distance Measures

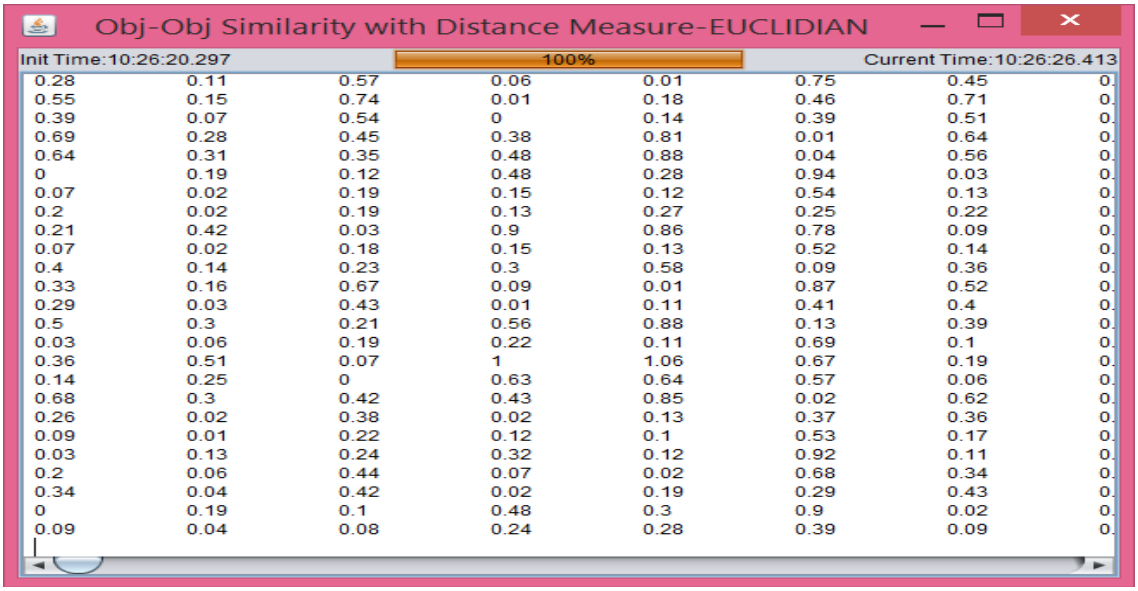

Fig: 5 Object Similarities with Distance Measure –Euclidian

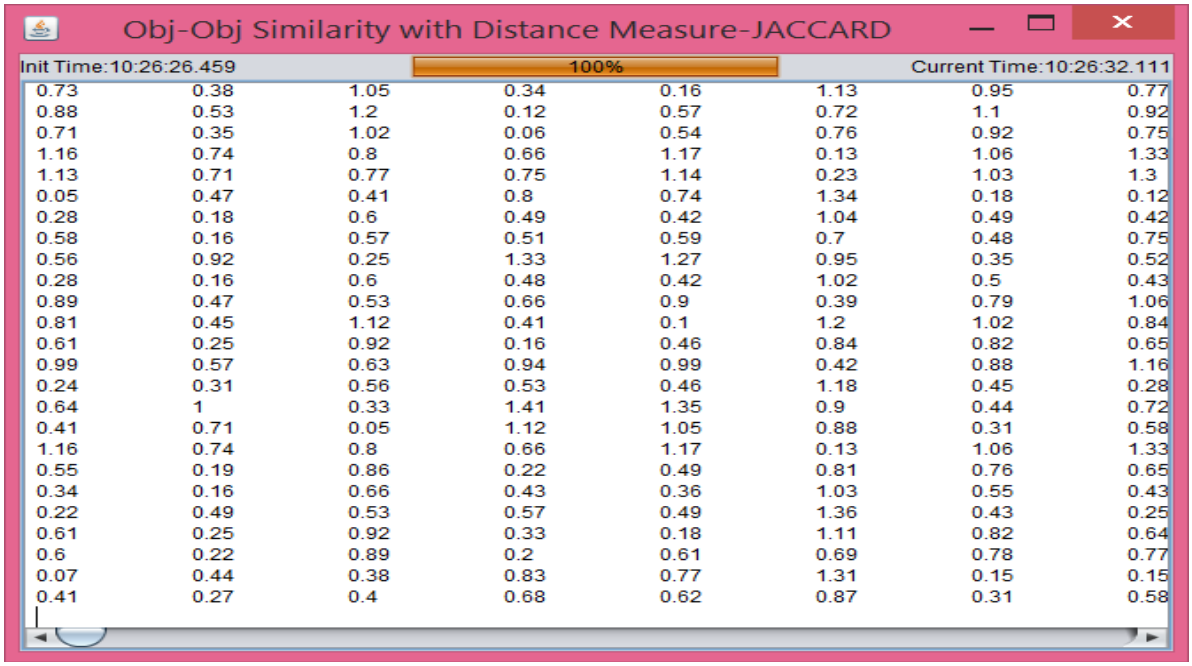

Fig: 6 Object Similarities with Distance Measure Jaccard

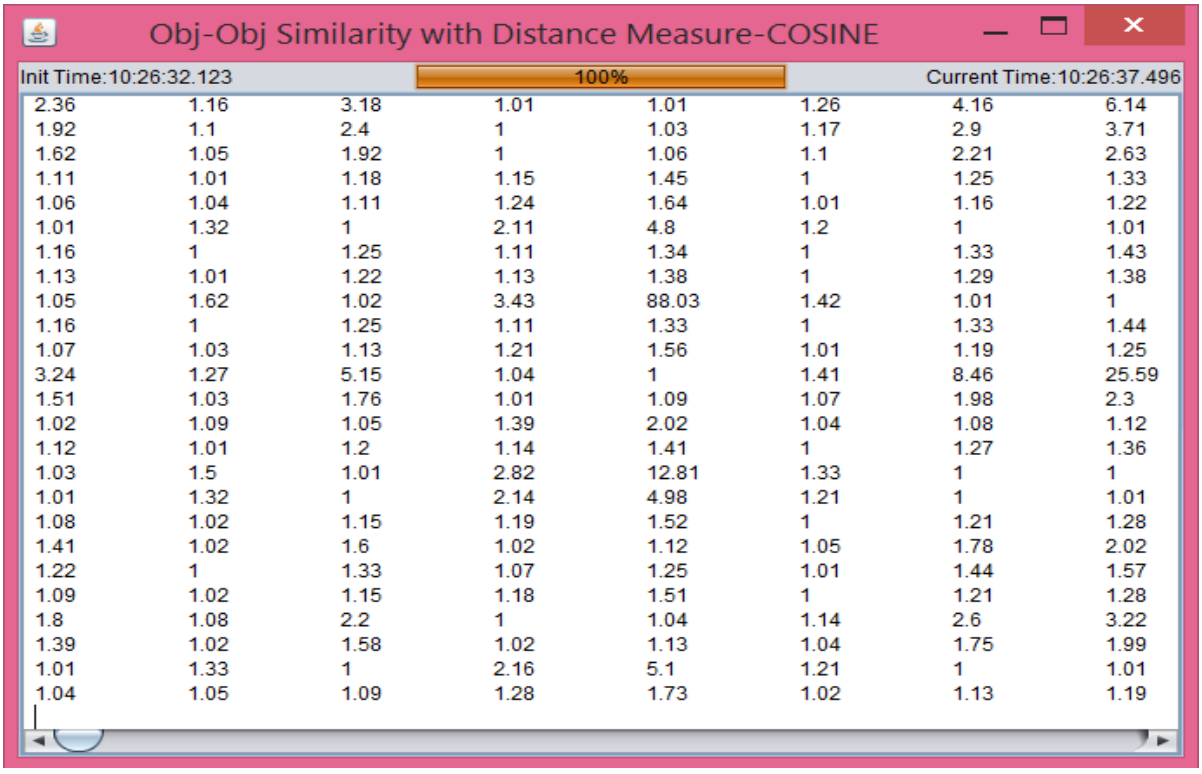

Fig: 7 Object Similarities with Distance Measure-Cosine

Fig: 8 Cluster Qualities

Fig: 9 Cluster Report

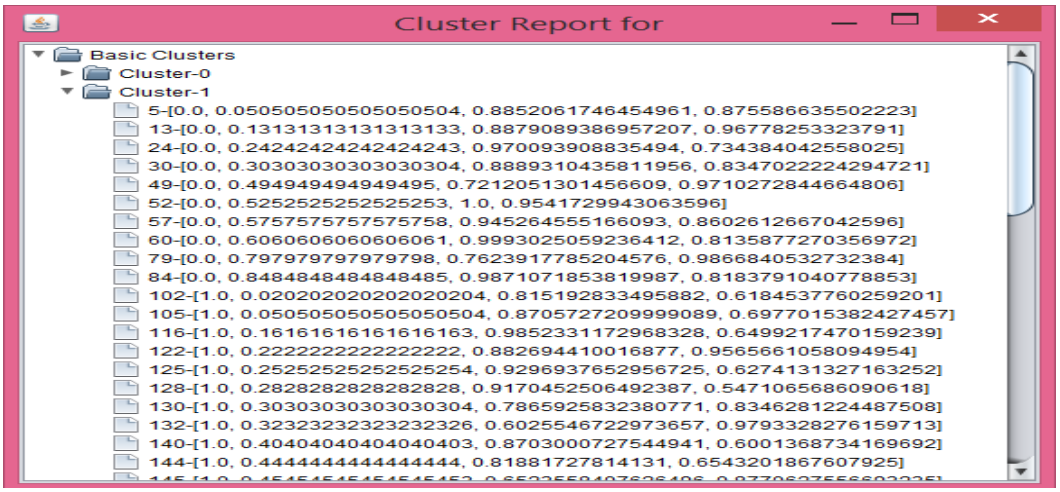

Fig: 9 Description of a Cluster

V. CONCLUSION

Hence proposed work is an ensemble of distance measures applied concurrently to effectively mine the data generated from wearable and implantable body sensor data. Simulation results shows how different distance measures are applied on supplied datasets. Cluster quality measure Davies Bouldin index is such that if increase in index value means less quality and decrease in value means high quality. It should not get zero because no quality. Proposed clustering algorithm is effective in terms of time complexity and adaptable for data stream clustering.

VI. FUTURE WORK

Proposed work has some manual intervention to tune the distance threshold parameters. To automate that evolutionary computing has to be used in future. Other provision of enhancement includes data stream clustering using proposed context aware distance similarity measure selection.

REFERENCES

[1] Kinsella, K.; Phillips, D.R. Global aging: The challenge of success. Pop. Bull. 2005, 60, 1-42.
[2] Sangita Singh (2013) “Integrated Analytics: The Way Forward in Healthcare.”
[3] I. F. Akyildiz, M. Pierobon, S. Balasubramaniam, and Y. Koucheryavy (2015), “The Internet of Bio-Nanothings”.
[4] Davies, David L.; Bouldin, Donald W. (1979)."A Cluster Separation Measure".IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2): 224–227. doi:10.1109/TPAMI.1979.4766909
[5] StepanIvanov, KritiBhargava, and William Donnelly, "Precision Farming: Sensor Analytics", IEEE Intelligent Systems, pp. 76-80 , 2015.
[6] HadiBanaee and Amy Loutfi,"Data-Driven Rule Mining and Representation of Temporal Patterns in Physiological Sensor Data",IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 19, NO. 5, pp. 1557-1566, SEPTEMBER 2015.
[7] GirmaKejela, RuiMaximoEsteves, and ChunmingRong, "Predictive Analytics of Sensor Data Using Distributed Machine Learning Techniques", IEEE 6th International Conference on Cloud Computing Technology and Science, pp. 626-631, 2014.
[8] Kluwer Academic Publishers ,"Managing And Mining Sensor Data" Edited by CHARU C. AGGARWAL IBM T. J. Watson Research Center, Yorktown Heights, NY, USA,

AUTHOR PROFILE

Name: Y.INDU

Designation: Asst.prof

Working At: Vardhaman College of Engineering

Shamshabad,Hyderabad

Name: V.Uma Maheswari

Designation: Asst.prof

Working At: Vardhaman College of Engineering

Shamshabad, Hyderabad