

ADAPTIVE REAL TIME DATA MINING METHODOLOGY FOR WIRELESS BODY AREA NETWORK BASED HEALTHCARE APPLICATIONS

Dipti Durgesh Patil ¹ and Vijay M. Wadhai ²

¹Department of Computer Engineering, MAEER's MITCOE, Pune, India
dipti.dpatil@yahoo.com

²Principal, MAEER's MITCOE, Pune, India
wadhai.vijay@gmail.com

ABSTRACT

Since the population is growing, the need for high quality and efficient healthcare, both at home and in hospital, is becoming more important. This paper presents the innovative wireless sensor network based Mobile Real-time Health care Monitoring (WMRHM) framework which has the capacity of giving health predictions online based on continuously monitored real time vital body signals. Developments in sensors, miniaturization of low-power microelectronics, and wireless networks are becoming a significant opportunity for improving the quality of health care services. Physiological signals like ECG, EEG, SpO2, BP etc. can be monitor through wireless sensor networks and analyzed with the help of data mining techniques. These real-time signals are continuous in nature and abruptly changing hence there is a need to apply an efficient and concept adapting real-time data stream mining techniques for taking intelligent health care decisions online. Because of the high speed and huge volume data set in data streams, the traditional classification technologies are no longer applicable. The most important criteria are to solve the real-time data streams mining problem with 'concept drift' efficiently. This paper presents the state-of-the art in this field with growing vitality and introduces the methods for detecting concept drift in data stream, then gives a significant summary of existing approaches to the problem of concept drift. The work is focused on applying these real time stream mining algorithms on vital signals of human body in Wireless Body Area Network(WBAN) based health care environment.

KEYWORDS

Wireless Sensor Networks, WBAN, Real-time data stream mining, concept-drift, physiological signal processing, Health Care,

1. INTRODUCTION

Research in data mining with respect to vital signal analysis is becoming very significant in order to provide good services to the patient's health. Different mobile monitoring systems [1], [2], [3] have been proposed to enhance people comfort, healthcare efficiency, and illness prevention [4]. While many efforts have been devoted to improve the architecture and the connectivity among devices [5], [6], less attention has been devoted to the development of analysis techniques to assess the current health status of monitored people. In particular, a fundamental task is the invention of adaptive and computationally efficient framework which will continuously monitor health of patient by analyzing its different physiological signals.

This paper presents an innovative wireless sensor network based Mobile Real-time Health care monitoring (WMRHM) framework which applies data mining techniques on real-time vital signals acquired through WSN and predict the health risk of the monitored person. The

prediction will be based on patient’s historical rule base, domain expert rules and currently monitored real-time signals. The proposed framework architecture of WMRHM is depicted in Figure 1.

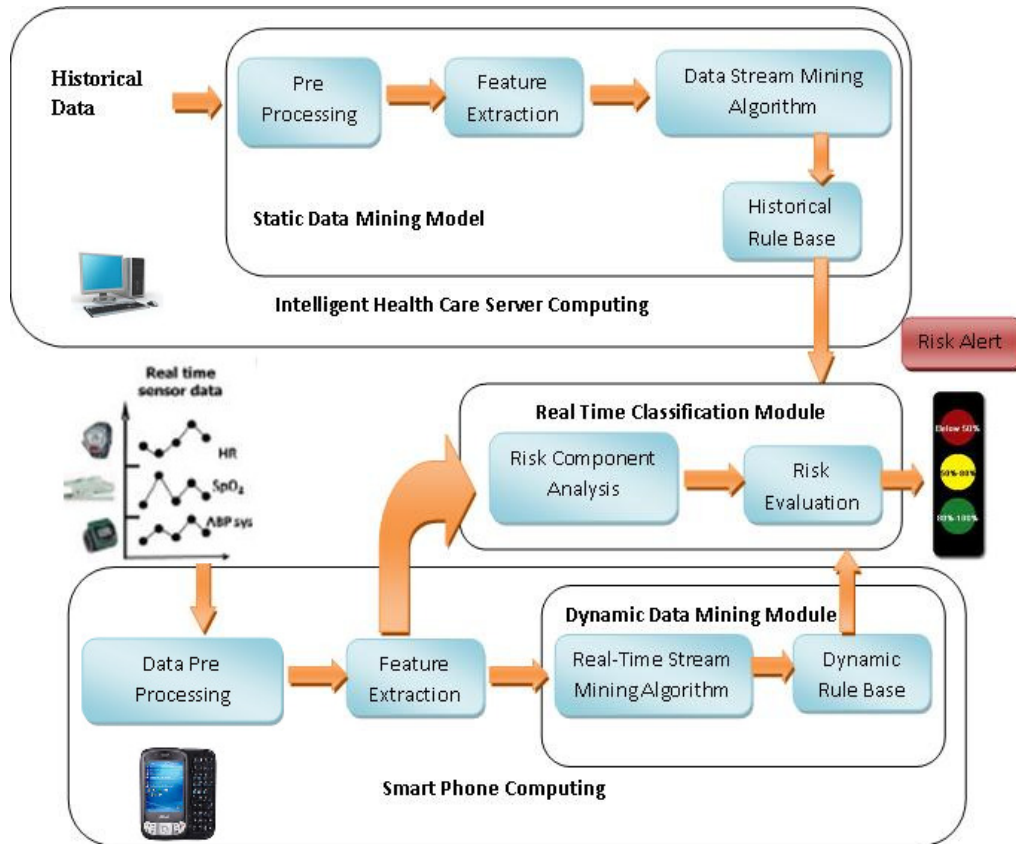


Figure 1: Wireless sensor network based Mobile Real-Time Health Monitoring System (W-MRHM) Framework

Data streams flow in and out from a computer system continuously and with varying update rates. They are temporally ordered, fast changing, massive, and potentially infinite.[7][17] It may be impossible to store an entire data stream or to scan through it multiple times due to its tremendous volume. So there is a need of analyzing this continuous data online without the overhead of storing it on a disk.

There exists a dynamic and promising field called data stream mining and knowledge discovery. To acquire knowledge base from raw data, emphasis is placed on innovative data stream mining concepts and techniques. This paper contains the general architecture of real-time data stream mining systems (RT-DSMS), different types of concept adapting algorithms, and finally finding useful patterns or knowledge from real-time data. Data streams are with the characteristics dynamic, non stationary, continuous, large volume, unstoppable, infinite.

The advanced research domain in DSM system is to handle concept drift in real-time data. While processing the data noise, errors, unwanted data, missing values have to be removed. There are many proposed classification algorithms for concept drifting data streams. These algorithms support multidimensional analysis and decision making. Additional data analysis techniques are required for in-depth analysis, characterization of data changes over time. In addition, huge volumes of data can be accumulated beyond databases and data warehouses. Fig

1 shows the general data stream mining process. In applications like video surveillance, weather forecasting, telecommunication, sensor networks, satellites, call records, vital signals monitoring; data stream mining plays a key role to analyze the continuous data. The effective and efficient analysis of this data in such different forms becomes a challenging task. Also the issue of memory constraints has to handle as enormous data is generated continuously.

Expert system technologies, which typically rely on users or domain experts manually, input knowledge into knowledge bases. Unfortunately, this procedure is prone to biases and errors, and is extremely time-consuming and costly. Real Time Data Stream Mining (RT-DSM) process performs data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research. Hence, advanced RT-DSM algorithms are discussed in this paper and how they can be applied on vital signals of human body for health care is depicted in figure 2.

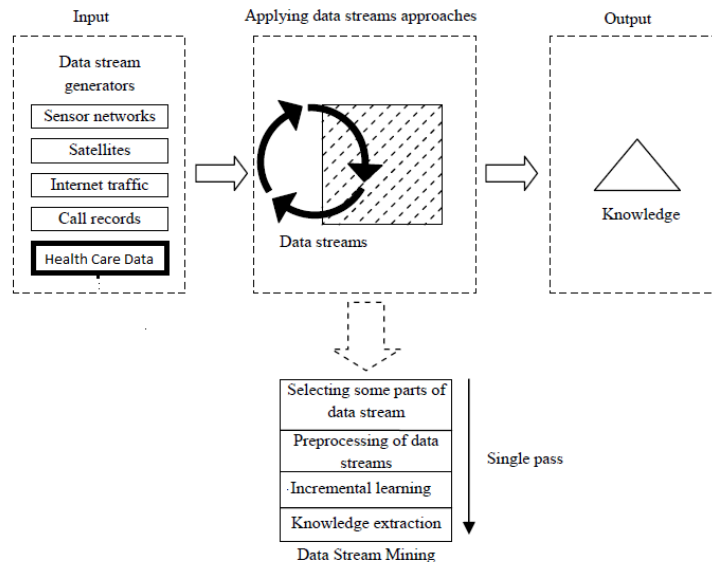


Figure 2: General Process of Data Stream Mining.

Section 2 discusses various issues in real-time data stream mining. Work related to Stream mining process and algorithms is described in section3. Section 4 explains algorithmic strategies for RT-DSM. The stress is on Ensemble-Based Classifiers. Section 5 depicts how these mining techniques can be applied on vital signals of human body for health care predictions and some of the results after applying data mining techniques on health care data in real-time. Section 6 represents Conclusion and future work.

2. Issues in Real Time Data Stream Mining

Issues and challenges beneath the concept drift are discussed below [7] [8]:

- A. **Robustness issue:** The noise problem is more crucial for stream data mining, because it is difficult to distinguish noise from changes caused by concept drift. If an algorithm is too eager to adapt to concept changes, it may over fit noise and might be interpreting it as data from a new concept. If an algorithm is too old fashioned and slow to adapt, it may overlook important changes.

- B. **Adaptation issue:** the concept generating a data stream drifts with time due to changes in the environment. These changes cause the model learned from old data is obsolete, and model updating is necessary.
- C. **Performance issue:** To assure on-line responses with limited resources, continuous mining should be “fast and light”, that is:
 - a. Learning should be done very fast, preferably in one pass of the data;
 - b. Algorithms should make light demands on memory resources.
- D. **Sampling data from a stream:** value or set of values at a point in time and/or space.
- E. **Filtering a data stream:** Extract only the specific data that you want to see, and then display it in the manner that you want to see it. To address these issues, analysis of distinct algorithms and strategies is required with modest resource consumption.

The core assumption when dealing with the concept drift problem is uncertainty about the future. If it is assumed that the source S_t of the target instance S_{t+1} is not known with certainty, it can be assumed, estimated or predicted but there is no certainty. Otherwise the data can be decomposed into two separate data sets and learned as individual models or in a combined manner.

3. RELATED WORK

Background theory of real-time data stream mining process is explained below:

- A. **Concept drift:** The underlying concept changes over time, so the learner should adapt to this change. It degrades the accuracy of classification system up to a point that the expected quality. Concept drift occurs during the classification mining process of data stream. Accuracy has been used to detect concept drift, which is sensitive to noise and affected by the effectiveness of the chosen classification algorithm.
- B. **Concept-evolution:** New classes evolve in the stream, which makes classification difficult. This change can be real or virtual. Concept drifts can be grouped into two main families: abrupt and gradual. The abrupt type refers to situations where changes can be modelled suddenly and gradual drift is step-like changes affecting the environment in which the classification system is deployed. Gradual models situations where the process slowly evolves over time.
- C. **Change Detection Algorithms:** Designing more efficient, accurate and parameter-free methods to detect change, maintain sets of examples and compute statistics. Trying to prove that the framework and the methods are useful, efficient and easy to use is tedious job. The imminent need for turning raw data into useful information and knowledge augments introduces development of systems, algorithms and frameworks that address streaming challenges. The storage, querying and mining of such data sets are highly computationally challenging tasks. Mining data streams is concerned with extracting knowledge structures represented in models and patterns in non stationary streams of information. Generally, two main challenges are designing fast mining methods for data streams and need to promptly detect changing concepts and data distribution because of highly dynamic nature of data streams.
- D. **Concept drift adaptation process:** Concept drift refers to the learning problem where the target concept to be predicted changes over time in some unforeseen behaviors. It is commonly found in many dynamic environments, such as data streams, P2P systems, etc. Real-world examples include network intrusion detection, spam detection, fraud

detection, epidemiological, and climate or demographic data, etc. Figure 3 depicts the incremental process of single learning instance [17], where the training model evolves with the concept drift and accordingly testing is carried out.

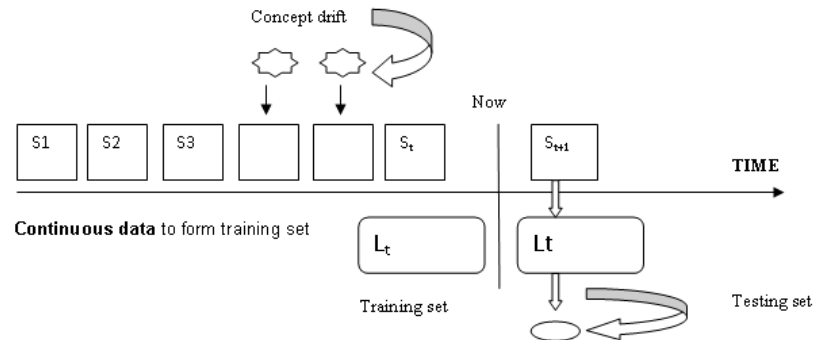


Figure 3: Incremental Learning of Single Learning Instance.

4. ALGORITHMIC STRATEGIES

Moving towards the innovative ideas, selecting the algorithm which is proper and efficient in giving results is very important.

Algorithms: General reasons for selecting the algorithms [9]:

- Popularity
- Flexibility
- Handling high dimensionality
- Applicability

There are different categories to adopt concept drifts.

Following categories include different algorithms.

1. Clustering
2. Decision Trees
3. Ensemble based classification

We are focusing on ensemble based classification algorithm. A classifier is said to predict randomly, if the probability of data point x being classified to a class c is equal to c 's class distribution in the current data block [10].

Algorithm 1: Ensemble Building

Input: Data stream with class labels available intermittently,

$\bar{\epsilon}$ = Permitted error

α = Classifier Precision

Output: A set of classifiers, Global set G ,

$G = \{C_1, C_2, C_3, \dots, C_n\}$

Classification of Testing data

1. Global set $\leftarrow G \{ \}$;
2. Ensemble set $\leftarrow G \{ \}$;
- $E \leftarrow \{ \}$;

Training

3. New Classifier Required \leftarrow true;
4. $E \leftarrow \{ \}$;
5. Get data chunk T from input stream with class label

6. MaxMSE \leftarrow classification error for data set T using a classifier predicting random
7. CE_i \leftarrow classification error for data set T using Classifier C_i
8. **for** classifier C_i in G
9. **if** CE_i < τ
10. New Classifier Required \leftarrow false;
11. GO TO *Training*
12. **endif**
13. **endfor**
14. **for** classifier C_i in G
15. **if** CE_i < (AcceptanceFactor * MaxMSE)
16. E \leftarrow {E} Union { C_i }
17. W_i \leftarrow MaxMSE - CE_i
18. **endif**
19. **endfor**
20. CE Ensemble \leftarrow Calculate Classification error using ensemble set E with weights
21. **if** (CE Ensemble < τ)
22. New Classifier Required \leftarrow false;
23. GO TO *Training*
24. **endif**
25. **if** (New Classifier Required = false)
26. C_i \leftarrow Build Classifier (Data Chunk T, Classifier Precision α)
27. G \leftarrow {G} Union {C_i}
28. GO TO *Training*
29. **endif**

End *Training*

Testing

30. Classify the incoming stream using ensemble set E until the next set of labeled data is available.

Algorithm 2: Training the Dynamical Discriminative Model

At the initial stage, the algorithm uses the *iterative reweighted least squares* method to train a logistic regression classifier based on an off-line data set. Then the parameter vector of the initial classifier is used as the initial value of w_t . Based on the same data set, the initial value of covariance of the measurement noise and parameter vector can also be estimated online. Because the covariance of w_t which denotes the degree of concept drift could change over time, we estimate its value based on a length-fixed buffer when it is filled with new examples and then update the related equations periodically. The performance of a classifier is quite stable when the size of the buffer is varied [11].

Input: S: a dataset from the incoming stream

C: a off-line dataset for evaluating the initial value of parameters

K: a size-fixed buffer for estimating a

Output: w_t : a series of parameter vector of classifier for each time stamp learn the initial parameter vector w_t from C using the IRLS method;

estimate the value of a from C using ;

while S not empty **do**

 get an instance x_t from S;

```

compute the prior estimate for  $w_t$  and  $P_t$ 
compute the posterior estimate for  $w_t$  and  $P_t$  using
output the posterior estimate of  $w_t$ ;
if the buffer  $K$  is full then
estimate the value of  $a$  from  $K$  using ;
 $P_t = aI$ ;
empty  $K$ ;
end

```

Algorithm 3: Adaptive Ensemble Classifier

The ensemble classifier are built and updated in an online manner. Once a new training instance arrives, it is used to update the ensemble classifier in an online bagging scheme. We store the new-comer instance in the predefined-size evaluation set. If the evaluation set is full, the least recent instance in the evaluation set will be removed. Whenever a chunk (the chunk size is smaller than the evaluation set size) of new instances arrive, we perform ensemble reconstruction and subset selection. The most recent chunk of training data is used to evaluate the ensemble for the current concept, and decide how much component classifiers should be dropped and which to be dropped. New classifiers are constructed and added to the ensemble to keep the ensemble size constant. Besides ensemble reconstruction, we choose a sub-optimal subset of the component classifiers that have the best accuracy in the evaluation set to participate in the final decision [12].

E: ensemble classifier N: Chunk-size
I: new instance M: Training set size ES: evaluation set

```

For (n = 1...M)
{
UpdateEnsemble (E, In);
UpdateEvaluationSet (In,ES);
If (n%N==0)
{
Reconstruct(E,ES);
SubsetSelection(E,ES)
}}

```

Algorithm 4: Online Bagging Algorithm

Bagging usually works better than the individual component classifier. We can expect better accuracy than a single classifier by using the bagging ensemble classifier. Compared with boosting, bagging is less affected by noise in the training data. It implies that bagging may work better than boosting in real world application. Modified bagging algorithm is given here: [13]

Inputs: ensemble E, Ensemble Size S, training example T

On-line learning algorithm for the ensemble members

OnlineBaseLearningAlg.

1. for t=1 to S do
2. $K \leftarrow$ Poisson
3. while $K > 0$ do
4. $hm =$ OnlineBaseLearningAlg(hm, T)
5. $K = K - 1$
6. end while
7. end for

Output: updated ensemble h.

Algorithm 5: DWCDs: Double Window Based Classification in real time data stream mining

Due to limited size of sliding window the number of samples consider are less and may not take into account the concept drift. To overcome this problem Double Window Based Classification in real time data stream mining (DWCDS) is proposed. It is based on the changes of the original data distribution in the window to detect concept drifts. Correspondingly, the window sizes are adjusted dynamically to enhance the adaptability to concept drifts. Experiments show that DWCDS performs better on the concept drift detection, the ability of robustness to noise and the accuracy of classification. [16]

Input: Training set DSTR

Test set: DSTE, Attribute set A, Maximum height of trees h_0 , Basic classifier count N, Capacity of each basic classifier K, Sliding Window SW, Minimum threshold of window MinSW, Maximum Threshold of window MaxSW, Basic window w, Minimum no. of split instances nmin, Coefficient of drift warning t_1 , Coefficient of drift t_2 ;

Output: Error rate of classification

Procedure: DWCDS { DSTR, DSTE, A, h_0 , N, K, SW, MinSW, MaxSW, w, nmin, t_1 , t_2 }

```

{
  For (i=1; i<(basic classifier count-N); i++)
    Generate k-random decision trees as a basic classifier CTi using the data in Sliding Window;
    While (a new instance arrives)
      {
        If (Sliding Window ==full)
          T: Detect concept drifts using the data distribution changes of current streaming data in the sliding window (SW)
          and a basic window (w);
          If (a concept drift occurs)
            T: 1. Delete the classifier CTi with the worst performance on classification from CT;
              2. Build a new basic classifier using the data that has a new concept in SW and put it into CT;
              3. Adjust the size of the sliding window;}}
  For (each test instance  $\in$  DSTE )
    Classify it in a voting mechanism using ensemble classifier {CTi}
  Return the error rate of classification;
}

```

The algorithms discussed above are applied and tested on different datasets. For verifying the applicability and flexibility of these algorithms the accuracy evaluation of these algorithms are given in Table1.

Table 1 Comparative analysis of RT-DSM algorithms

Sr. No.	Name of Algorithm	Accuracy	Dataset used
1.	Ensemble Building	95%	Nursery data
2.	Training the Dynamical Discriminative Model	93.45%	Synthetic data
3.	Adaptive Ensemble Classifier	97.23%	SEA dataset
4.	Online Bagging Algorithm:	More than 96%	Hyper-plane Dataset
5.	DWCDS	93.5 to 98%	KDDCup99 Dataset

5. IMPLEMENTATION AND RESULTS

An objective of a health process is one where patients can stay healthy with the support of expert medical advice when they need it, at any location and any time. An associated aim would be the development of a system which places increased emphasis on preventative measures as a first point of contact with the patient. As the vital signals plays key role for predicting health status of a human, and these signals are continuous in nature(ECG, EEG, Heart Rate, SPO2 etc.), we propose to apply RT-DSM on these signals. Some of the basic data mining algorithms like K-means and Graph Theoretic algorithm are applied on the vital signals like ABPdias(diastolic blood pressure), ABPsys(sysstolic), Heart rate, SPO2.

Publicly available clinical data have been selected to validate the effectiveness of the proposed framework. Focusing on the intensive care scenario, the multi-parameter intelligent monitoring for intensive care (MIMIC) database [19] contains nearly 200 patient days of real-time signals. We used 64 records from the numeric section of the database, which provides measures sampled at 1 Hz. For each record, the gender, age, and disease of the patient are known. Same signals can be taken in real-time with the help of wireless body sensors and behavior of these signals can be learned on the fly to predict the patient’s health status with the help of above discussed ensemble-based classifiers.

The sample results are shown in figure 4 for SPO2 signal and for ABP (dias) signal in figure 5 with K-means algorithm. The clustering algorithms applied are offline in nature, so they will be used for formation of historical rule base. We propose to apply above discussed real-time data stream mining algorithms on the vital signals to dynamically update the rule base and predict the health risk accurately.

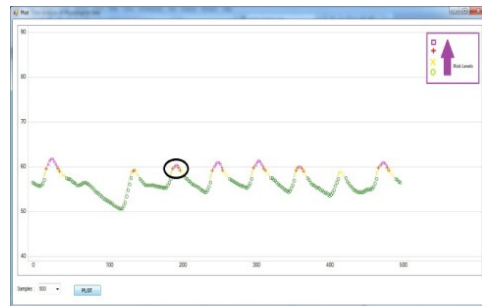


Figure 4: SPO2 Signal to Analyze Risk Level.

This will help to take immediate preventive actions in case high health risk. Also there will not be any need to keep patient in ICU in wired environment, instead patient can be continuously monitored from his home and alerted in a risky situations.

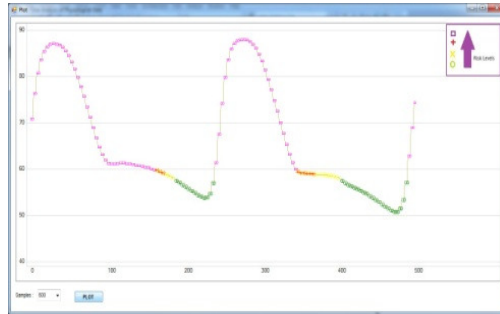


Figure 5: ABP[dias] Signal to analyze Risk Level.

6. CONCLUSION AND FUTURE WORK

An innovative WMRHM framework has been discussed. The framework is innovative as it dynamically adapts to the changes happened in the vital signals and updates the model for health risk predictions. The probability of accurate predictions is high enough as it considers Historical rule base, domain expert's rule and real-time rule model for analyzing the health status. Different ensemble based classifier systems for RT-DSM has been discussed. All these methods are capable of performing any-time classification, learning in one scan and detecting drift in the underlying concept. The important issue of adapting concept drifts has been solved. We are focusing on real time data stream mining of health care data. The experiments are carried out on various vital signals of human body by applying the data mining algorithms like K-means and Graph Theoretic algorithm to predict the health risk level. The same data will be taken in real-time and dynamic algorithms will be applied on these vital signals to continuously monitor the health status. Providing mobility to patients through wireless networks and helping doctors to take preventive actions immediately by assisting them through real time decision support system is the motivation of this work.

REFERENCES

- [1] H.-C. Wu, C.-H. Lin, K.-C. Wang, S.-C. Wang, C.-H. Chen, S.-T. Young and T.-S. Kuo, "A mobile system for real-time patient-monitoring with integrated physiological signal processing," in Proc. 1st Joint BMES/Eng. Med. Biol. Soc. Conf., Atlanta, GA, 1999, p. 712.
- [2] Y.-H. Lin, I.-C. Jan, P. Chow-InKo, Y.-Y. Chen, J.-M. Wong, and G.-J. Jan, "A wireless PDA-based physiological monitoring system for patient transport," IEEE Trans. Inf. Technol. Biomed., vol. 8, no. 4, pp. 439–447, Dec. 2004.
- [3] R.-G. Lee, K.-C. Chen, C.-C. Hsiao and C.-L. Tseng, "A mobile care system with alert mechanism," IEEE Trans. Inf. Technol. Biomed., vol. 11, no. 5, pp. 507–517, Sep. 2007.
- [4] N. Saranummi, "Information technology in biomedicine," IEEE Trans. Biomed. Eng., vol. 49, no. 12, pp. 1385–1386, Dec. 2002.
- [5] K. Lorincz, D. J. Malan, T. R. F. Fulford-Jones, A. Nawoj, A. Clavel, V. Shnayder, G. Mainland, M. Welsh, and S. Moulton, "Sensor networks for emergency response: Challenges and opportunities," IEEE Pervasive Comput., vol. 3, no. 4, pp. 16–23, Oct. 2004.
- [6] B. G. Cellar, N. H. Lovell and J. Basilakis, "Using information technology to improve the management of chronic disease. Med. J. of Australia, 2003; 179 (5): 242-246.

- [7] P. Domingos and G. Hulten. "Mining high-speed data Streams" In Knowledge Discovery and Data Mining, 2000, pp 71–80.
- [8] M. B. Harries, C. Sammut, and K. Horn. "Extracting hidden context. Machine Learning", 32(2): 1998. pp 101–126
- [9] G. Hulten, L.Spencer, and P. Domingos "Mining Time changing data streams", international conference on Knowledge discovery and data mining KDD 01 vol. 1, Publisher: ACM Press 2001, pp 97-106
- [10] Sasthakumar Ramamurthy, Raj Bhatnagar "Tracking Recurrent Concept Drift in Streaming data using Ensemble Classifiers" Machine Learning and Applications, in Proc.ICMLA 2007, pp 404 - 409
- [11] Bai Su, Yi-Dong Shen Wei Xu "Modeling Concept Drift from The Perspective of Classifiers" Cybernetics and Intelligent Systems, in proc. IEEE, 2008, pp 1055 - 1060
- [12] Dengyuan Wu1,Ying Liu, Zhendong Mao, Weishan Ma1,Tao He, " An Adaptive Ensemble Classifier For Concept Drifting Stream" , Computational Intelligence and Data Mining CIDM '09. IEEE Symposium APRI 2009, pp 69 – 75
- [13] Leandro L. Minku, Allan P. White, and Xin Yao, "The impact of diversity on learning in the presence of concept drift" IEEE Transactions on Knowledge and Data Engineering MAY 2010 pp 730 - 742
- [14] Micheline Kamber " Data Mining: Concepts and Techniques", Second Edition Jiawei Han University of Illinois at Urbana-Champaign ISBN 13: 978-1-55860-901-3 ISBN 10: 1-55860-901-6, 2006
- [15] Albert Bifet, Geoff Holmes, Richard Kirkby and Bernhard Pfahringer "Data Stream Mining a Practiccal Approach" MAY 2011.
- [16] Qun Zhu; Xuegang Hu; Yuhong Zhang; Peipei Li; Xindong Wu; " A Double-Window-Based Classification Algorithm for Concept Drifting Data Streams", In Granular Computing (GrC), IEEE International Conference on 14-16 Aug. 2010 pp 639 – 644
- [17] Indrè Zlibaite,"Learning under Concept Drift: an Overview" Overview", Technical report, Vilnius University, 2009 techniques, related areas, applications Subjects: Artificial Intelligence arXiv:1010.4784v1 [cs.AI],OCT 2010 Report number: 2009
- [18] Mahnoosh Kholghi, "An Analytical framework for data stream mining techniques based on challenges and requirements", IJEST, March 2011
- [19] The MIMIC database on PhysioBank 2007, Oct. [Online]. <http://www.physionet.org/physiobank/database/mimicdb>
- [20] Daniele Apiletti, Elena Baralis, Member, IEEE, Giulia Bruno, and Tania Cerquitelli, "Real-Time Analysis of Physiological Data to Support Medical Applications", IEEE transactions on information technology in biomedicine, vol. 13, NO. 3, MAY 2009, pp 313-321

Authors

Dipti D. Patil has received B.E. degree in Computer Engineering from Mumbai University in 2002 and M.E. degree in Computer Engineering from Mumbai University, India in 2008. She is pursuing her PhD from Sant Gadgebaba Amravati University. She is currently working as Assistant Professor in MITCOE, Pune. She has filed patents in data mining area and published various numbers of international papers. Her Research interests include machine learning, Medical Signal Processing and Wireless Body Area Networks.



Dr. Vijay M.Wadhai received his B.E. from Nagpur University in 1986, M.E. from Gulbarga University in 1995 and Ph.D. degree from Amravati University in 2007. He has experience of 26 years which includes both academic (19 years) and research (7 years). He has been working as the Principal of MAEER's MITCOE, Pune since 2010. He had worked as Dean of Research, MITSOT, MAE, Pune (from 2009 to 2010) and handled the post of Director - Research and Development, Intelligent Radio Frequency (IRF) Group, Pune. He is currently guiding various numbers of students for their PhD work in both Computers and Electronics & Telecommunication area. His research interest includes Deductive Databases, Knowledge Discovery and Data Mining, Data Security, Natural Language processing, Cognitive Radio and Wireless Communication, Spectrum Management, Wireless Sensor Network, ASIC Design - VLSI, Advance Network Design. He is a member of LMISTE, MIETE, MIEEE, MIES and GISFI (Member Convergence Group), India. He has filed numbers of patents in his research areas and also published quality International journal papers.

