Data Mining Approach on Human Activity Pattern Mining To Determine The Health Risk Factor Using Smart Device

EstherShilpa D [1], T.H Ferioz Khan [2]
[1, 2] Dept of CSE
Assistant Professor (SG), Dept of CSE
[1, 2] SRM IST, Ramapuram, Chennai-600095

Abstract- *In this Paper, the huge data growth in biomedical and healthcare businesses, accurate analysis of medical data benefits from early detection, patient care, and community services. However, the analysis accuracy is reduced if the quality of the medical data is incomplete. In addition, the healthcare industry collects large amounts of healthcare information which cannot be mined to find unknown information for efficient evaluation. Discovery of buried patterns frequently goes unexploited. We propose frequent mining and prediction model to measure and analyse energy usage changes sparked by occupants' behaviour. The data from smart meters are recursively mined in the quantum/data slice of 24 hours, and the results are maintained across successive mining exercises. We also utilize the Bayesian network, a probabilistic graphical model, to predict the energy levels of humans. The proposed model is capable of short-term predictions ranging from next hour up to 24 hours and long-term prediction for days, weeks, months, or seasons. For the evaluation of the proposed mechanism, this research uses the UK Domestic Appliance Level Electricity dataset (UK-Dale) .*

Keywords- Healthcare Businesses, Machine Learning Algorithms, Predictive Models, Neuronal Network-based Multimodal Disease Risk Prediction

I. INTRODUCTION

Big data is the data that exceed the processing prospective of traditional database systems. The data is too big, too fast, or does not fit the limitations of traditional database architectures. The salient features of Big Data are Volume, Velocity, Variety, Value and Veracity.

Big Data plays an important role in healthcare. In healthcare, big data refers to electronic health records that are too big to manage with traditional software or with common data management tools. Big data in healthcare is overwhelming, not only because of its volume, but also because of the variety of data types and the speed with which it has to be managed. Healthcare analytics finds insights from unstructured, complex and noisy health records of patients to make better health decisions.

Described as a herd of development for health and disease, the study has successfully demonstrated the importance of individuals' developmental records in predicting and / or explaining the illnesses a person is suffering. In this current, large paper-based health data world, invaluable data are often unavailable at the right time in the hands of clinical care providers to allow better care. This is largely due to the inefficiencies of the paper-based system. Supervised learning is the machine learning task of disrupting a function of labelled training data. In this project we use supervised learning for the training. The ensemble of classifiers is a combination of several classifiers called base classifiers. Ensembles usually perform better than any single classifier. To create a good base classifier, the base classifiers must also be different, which means that for the same instance, the base classifiers should return different outputs and their errors should be in different instances.

The disease risk model is obtained through the combination of structured and unstructured features. Through the experiment, we draw a conclusion that the performance of our system model is better than other existing methods. The aim of this study is to predict whether a patient according to his activity belongs to the risk group. Formally, we consider the risk prediction model as the supervised learning method of machine learning. Use the patient's structured data to predict if the patient is at high risk. Machine learning techniques are widely used in many scientific fields, but their use in the medical literature is limited in part because of technical difficulties. K-mean clustering is a simple method of machine learning. The article introduces some basic ideas that underlie the k means algorithm, and then focuses on performing k-means Incremental modelling.

II. RELATED WORK

Lately, there has been a growing interest in using smart home technologies for detecting human activity patterns for health monitoring applications. The main goal is to learn occupant's behavioural characteristics as an approach to understand and predict their activities that could indicate health issues. In this section, we review existing work in the literature, which employ smart homes data to analyse users' behaviour.

Detecting human activities in smart homes by means of analyzing smart meters data is studied in [10]. The paper proposes two approaches to analyze and detect user's rou-tines. One approach uses Semi-Markov-Model (SMM) for data training and detecting individual habits and the other approach introduces impulse based method to detect Activity in Daily Living (ADL) which focuses on temporal analysis of activities that happen simultaneously. Similarly, the work in [11] proposes human activity detection for wellness monitoring of elderly people using classification of sensors related to the main activities in the smart home. Smart meters data are also used in [4] for activity recognition using Non-intrusive Appliance Load Monitoring (NALM) and Dempster-Shafer (D-S) theory of evidence. The study collects pre-processed data from homes to determine the electrical appliance usage patterns and then employs machine learning-based algorithm to isolate the major activities inside the home. The issue is that the study has to perform two steps on the data to completely isolate the main activities. Exploiting appliance usage pat-terns and identify them for sudden behavioural change is presented in [12]. The aim of the study is to provide around the clock monitoring system to support people's suffering from Alzheimer or Parkinson disease at minimum intrusion level. The study uses classification techniques to detect abnormal behavior of personal energy usage patterns in the home. Other studies such as [13] [15] and [16] although do not utilize smart meters data, they use Internet of Things (IOT) infrastructures in smart cities for developing applications that monitor and provide health services for patients.

Using data analytics for smart meters to detect and predict behavioural abnormality for remote health monitoring is discussed in [17] and [18]. Alam et al. [17] use everyday appliances usage from smart meter and smart plug data to trace regular activities and learn unique time segment groups of appliance's energy consumption. The study employs hierarchical probabilistic model-based detection to infer about discovered anomalous behavior. This in term can be used to understand the criticality of some abnormal behaviors for sustaining better health care. In [18] an experimental demonstration for observing and measuring energy consumption of appliances is presented. The study aims to provide a portrait of activities of daily living for elderly patients independently living at home. The data is also used to mine important patterns of changes for short-term and long-term anomaly detection of urgent health conditions. The work in [19], uses Bayesian networks to predict occupant behaviour from collected smart meters data. The study proposes behaviour as a service based on a single appliance, but does not pro-vide a model to be applied for real-world scenarios. Authors in [20] and [21], used time-series multi-label classifier to forecast appliance usage based on decision tree correlations, however, the study takes only the last 24-hour window along with appliance sequential relationships. Chelmis et al. [22] suggest a clustering approach to identify the distribution of consumers' temporal consumption patterns; however, the study does not consider appliance level usage details. This might not be applicable for human activity recognition since specific activities require individual and multiple appliance to appliance and time associations. The work in [23] considers the appliances' ON and OFF status to detect usage pattern using hierarchical and c-means clustering. However, the study does not consider the duration of appliance usage or the expected variations in the sequence of appliance usage. The work in [24] proposes graphical model based algorithm to predict human behavior and appliance interdependency patterns and use it to predict multiple appliance usages using a Bayesian model.

The above-discussed approaches do not consider appliance level usage patterns, which is critical in determining human activity variations. Furthermore, our experiments are conducted using a much larger dataset than existing studies although there are similarities in data analytics techniques between the proposed study and existing work

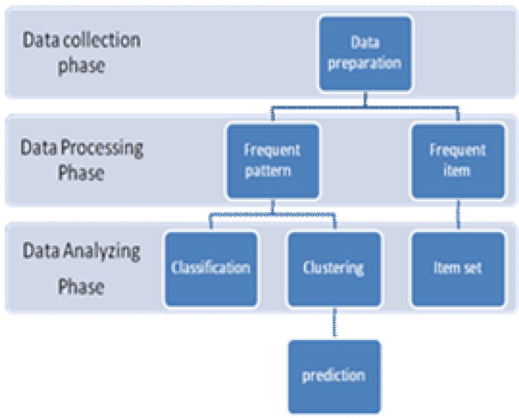


Figure 1: System Architecture

III. PROPOSED SYSTEM

Figure 1 represents the proposed model. It starts by cleaning and preparing the data and then applying frequent pattern for activity discovery. Based on what we develop a model for The Heterogeneity Human Activity Recognition from Smartphones and Smart watches.
We presented incremental frequent mining and prediction model based on Bayesian network. In our current work, through experiments

The Heterogeneity Dataset for Human Activity Recognition from Smartphone and Smart watch sensors consists of two datasets devised to investigate sensor heterogeneities' impacts on human activity recognition algorithms

The output of the system is utilized by user to know the energy level health care applications depending on the intended use. For example, a health care provider might only interested in knowing activities related to cognitive impairment where tracking the sequence of daily activities is crucial for reminding the patient when abnormal behavior is detected. Next subsection explain such processes and brie y outlines the theoretical background.

IV. HUMAN ACTIVITY RECOGNITION USING SMARTPHONES DATA SET

Description: Determine what activity a person is engaging in (e.g., WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) from data recorded by a smartphone (Samsung Galaxy S II) on the subject's waist. Using its embedded accelerometer and gyroscope, the data includes 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz.

•Size Training set: 7352 time steps with 561 features at each time step.

Classification, Clustering. Human Activity Recognition database built from the recordings of 30 subjects performing activities of daily living (ADL) while carrying a waist-mounted smartphone with embedded inertial sensors. The paper describes the use of an SVM on this data set, classifying each time step into one of the activities without taking temporal structure into account. This is a fairly naive method, and it could be interesting to compare this to a time series model like an HMM. Alternatively you could look at different classification methods or different ways of handling variability between subjects.

V. MINING ACTIVITY PATTERNS

Our goal is to develop a method that can automatically discover resident activity patterns over time from streaming sensor data. Even if the patterns are somehow discontinuous or have different event orders across their instances. Both situations happen quite frequently while dealing with human activity data. For example, consider the “meal preparation activity”. Most people will not perform this activity in exactly the same way each time, rather some of the steps or the order of the steps might be changed (variations). In addition the activity might be interrupted by irrelevant events such as answering the phone (discontinuous). The IFG method proposed in [11] finds such patterns in a static dataset. For example, the pattern a, b can be discovered from instances {b, x, c, a}, {a, b, q}, and {a, u, b}, despite the fact that the events are discontinuous and have varied orders.

We discover the frequent patterns from the current data batch $B\tau$ by using an extended version of FP Growth that is able to find patterns in streaming data and is also able to deal with varying frequencies across different regions of a physical space. After finding patterns in current data batch $B\tau$, we will update the time windows, and will prune any pattern that seems to be unpromising.

To find patterns in data, first a reduced batch $B\tau r$ is created from the current data batch $B\tau$. The reduced batch contains only frequent and sub frequent sensor events, which will be used for constructing longer patterns. A minimum support is required to identify such frequent and sub frequent events. IFPG uses as global minimum support, and it only identifies the frequent events. Here, we introduce the maximum sensor support error s to allow for the sub frequent patterns to be also discovered. We will also automatically derive multiple minimum supports values corresponding to different regions of the space. In mining real life activity patterns, the frequency of sensor events can vary across different regions of the home or other space. If the differences in sensor event frequencies across different regions of the space are not taken into account, the patterns that occur in less frequently used areas of the space might be ignored. For example, if the resident spends most of his/her time in the living-room during the day and only goes to the bedroom for sleeping, then the sensors will be triggered more frequently in the living-room than in the bedroom. Therefore when looking for frequent patterns, the sensor events in the bedroom might be ignored and consequently the sleep pattern might not be discovered. The same problem happens with different types of sensors, as usually the motion sensors are triggered much more frequently than other type of sensors such as cabinet sensors. This problem is known as “rare item” problem in market basket analysis and is usually addressed by providing multiple minimum support values [32]. Our proposed solution for solving the problem of rare items in data stream can be applied to other application domains, such as Web click mining. For example, a web page might have a lower chance of being visited by visitors, but we still might be interested in finding click patterns in such pages.

We will automatically derive multiple minimum sensor support values across space and over time. To do this, we identify different regions of the space using location tags l, corresponding to the functional areas such as bedroom, bathroom, etc. Also different types of sensor might exhibit varying frequencies. In our experiments, we categorized the sensors into motion sensors and interaction-based sensors. The motion sensors are those sensors tracking the motion of a person in a home, e.g. infra-red motion sensors. Interaction-based sensors, as we will call them “key sensors”, are the non-motion tracking sensors, such as cabinet sensors or RFID tags on items. Based on observing and analysing sensor frequencies in multiple smart homes, we found that a motion sensor might have a higher chance of being triggered than a key sensor in some regions. Hence we will derive separate minimum sensor supports for different sensor categories.

For the current data batch Bτ , we compute the minimum regional support for different categories of sensors as in Equation 3. Here l refers to a specific location. c refers to the sensor’s category, and Sc refers to the set of sensors in a pattern matches the general pattern, but does not exactly match any of the variations, it is added as a new variation. Otherwise it will be considered as a new general pattern.

Algorithm: Incremental Frequent Pattern Mining

Require: Transaction database (DB), Frequent pattern discovered database (FP_DB)

Ensure: Incremental discovery of frequent patterns, stored in frequent patterns discovered database (FP_DB)

1: For all Transaction data slice db24 in quanta of 24 hours in database DB do {Data is processed in slices of 24 hour period}
2: Determine database size Database_Sizedb24 for data slice/quantum db24
3: Mine Frequent patterns in FP_DBdb24 using extended FP-growth approach
4: For all Frequent Pattern FP in FP_DBdb24 do
5: Search a frequent pattern FP in FP_DB
6: If Frequent Pattern found then
7: Update frequent pattern in FP_DB
8: else
9: Add a new Frequent Pattern to FP_DB
10: End if
11: End for
12: For all Frequent Patterns in Database FP_DB increment Database Size by Database_Sizedb24
13: end for

At the end of each pattern the growth iteration, infrequent or highly discontinuous patterns and variations will be discarded as uninteresting patterns. Instead of solely using a pattern’s frequency as a measure of interest, we use a compression objective based on the minimum description length (MDL)

Using a compression objective allows us to take into account the ability of the pattern to compress a dataset with respect to pattern’s length and continuity

We continue extending the patterns by prefix and suffix at each iteration until no more interesting patterns are found. A post-processing step records attributes of the patterns, such as event durations and start times. We refer to the pruning process performed during the pattern growth on the current data batch as normal pruning. Note that it’s different from the tail pruning process which is performed on the time window to discard the unpromising patterns over time.

VI. CLUSTERING

K is an input to the algorithm for predictive analysis; it stands for the number of groupings that the algorithm must extract from a dataset, expressed algebraically as k. A K-means algorithm divides a given dataset into k clusters. The algorithm performs the following operations:

1. Pick k random items from the dataset and label them as cluster representatives.
2. Associate each remaining item in the dataset with the nearest cluster representative, using a Euclidean distance calculated by a similarity function.
3. Recalculate the new clusters’ representatives.
4. Repeat Steps 2 and 3 until the clusters do not change.

A representative of a cluster is the mathematical mean (average) of all items that belong to the same cluster.

VII. CLASSIFICATION

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods

To demonstrate the concept of Naïve Bayes Classification, consider the example displayed in the illustration above. As indicated, the objects can be classified as either GREEN or RED. Our task is to classify new cases as they arrive, i.e., decide to which class label they belong, based on the currently exiting objects.

Since there are twice as many GREEN objects as RED, it is reasonable to believe that a new case (which hasn't been observed yet) is twice as likely to have membership GREEN rather than RED. In the Bayesian analysis, this belief is known as the prior probability. Prior probabilities are based on previous experience, in this case the percentage of GREEN and RED objects, and often used to predict outcomes before they actually happen.

Thus, we can write:

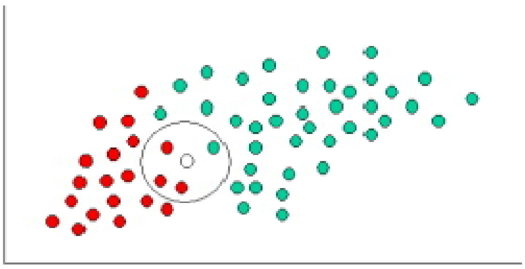
$$\text{Prior probability for GREEN} \propto \frac{\text{Number of GREEN objects}}{\text{Total number of objects}}$$

$$\text{Prior probability for RED} \propto \frac{\text{Number of RED objects}}{\text{Total number of objects}}$$

Since there is a total of 60 objects, 40 of which are GREEN and 20 RED, our prior probabilities for class membership are:

$$\text{Prior probability for GREEN} \propto \frac{40}{60}$$

$$\text{Prior probability for RED} \propto \frac{20}{60}$$



Having formulated our prior probability, we are now ready to classify a new object (WHITE circle). Since the objects are well clustered, it is reasonable to assume that the more GREEN (or RED) objects in the vicinity of X, the more likely that the new cases belong to that particular color. To measure this likelihood, we draw a circle around X which encompasses a number (to be chosen a priori) of points irrespective of their class labels. Then we calculate the number of points in the circle belonging to each class label. From this we calculate the likelihood:

$$\text{Likelihood of } X \text{ given GREEN} \propto \frac{\text{Number of GREEN in the vicinity of } X}{\text{Total number of GREEN cases}}$$

$$\text{Likelihood of } X \text{ given RED} \propto \frac{\text{Number of RED in the vicinity of } X}{\text{Total number of RED cases}}$$

From the illustration above, it is clear that Likelihood of X given GREEN is smaller than Likelihood of X given RED, since the circle encompasses 1 GREEN object and 3 RED ones. Thus:

$$\text{Probability of } X \text{ given GREEN} \propto \frac{1}{40}$$

$$\text{Probability of } X \text{ given RED} \propto \frac{3}{20}$$

Although the prior probabilities indicate that X may belong to GREEN (given that there are twice as many GREEN compared to RED) the likelihood indicates otherwise; that the class membership of X is RED (given that there are more RED objects in the vicinity of X than GREEN). In the Bayesian analysis, the final classification is produced by combining both sources of information, i.e., the prior and the likelihood, to form a posterior probability using the so-called Bayes' rule (named after Rev. Thomas Bayes 1702-1761).

$$\text{Posterior probability of } X \text{ being GREEN} \propto$$
$$\text{Prior probability of GREEN} \times \text{Likelihood of } X \text{ given GREEN}$$
$$= \frac{4}{6} \times \frac{1}{40} = \frac{1}{60}$$

$$\text{Posterior probability of } X \text{ being RED} \propto$$
$$\text{Prior probability of RED} \times \text{Likelihood of } X \text{ given RED}$$
$$= \frac{2}{6} \times \frac{3}{20} = \frac{1}{20}$$

Finally, we classify X as RED since its class membership achieves the largest posterior probability.

VIII. EXPERIENTAL RESULTS

For the evaluation of the proposed model, we performed our experiments using the dataset UK-Dale [9] along with the synthetic dataset to inspect intermediate and nal results.

We evaluated the performance of the following classifiers, all available in the Weka toolkit: Multilayer Perceptron,Random Forest, LMT, SVM, Simple Logistic and LogitBoost. Classifiers were trained and tested using a 10-foldcross validation method on the set of extracted features. The summary results for our activity recognition experiments are presented in datasets: phone in hand position and phone in pocket position. Overall, Multilayer Perceptron offered the highest performance, yielding89.48% accuracy for in-hand position and 89.72% accuracy for in-pocket position. SVM was the second most accurate for in-hand position. Its good performance has been supported by prior research on human activity recognition tasks. However, Our results also showed that Random Forest demonstrated high accuracy for both cases. Ta indicates that carrying the phone in pocket or in hand produced similar results. Hence, our recognition method is robust to smartphone position.

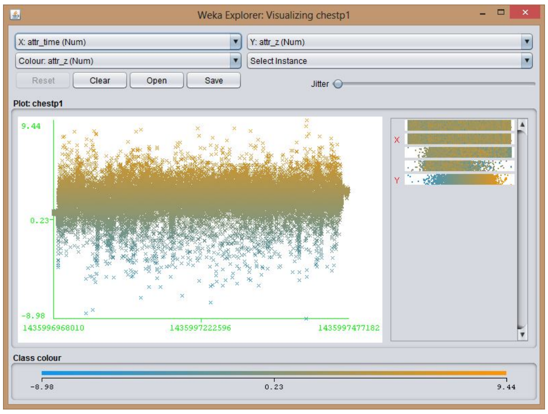


Figure 2: Clustering with k means

Fig 3 show the execution time for a person involved in activity of sitting, standing, jumping

Our recognition accuracy of up to 91.15% on various everyday activities using a single triaxial accelerometer was obtained. The data were acquired from multiple subjects under real-world conditions for two most common phone positions: smartphone in hand and smartphone in pants pocket. A new set of features was taken into account and different classifiers were used for evaluating recognition performance. Combining the three best classifiers using the average of probabilities method turned out to be the best classifier for activity recognition, outperforming all individual classifiers

IX. CONCLUSION

In this paper, we presented a model for recognizing human activities patterns from low resolution smart meters data. Occupants' habits and behavior follow a pattern that could be used in health applications to track the wellbeing of individuals living alone or those with self-limiting conditions. Most of these activities can be learned from appliance-to-appliance and appliance-to-time associations. We presented incremental frequent mining and prediction model based on Bayesian network. In our current work, through experiments, we found that 24-hour period was optimal for data mining, but we built the model to operate on any quantum of time. From the experiment results we have demonstrated the applicability of the proposed model to correctly detect multiple appliance usage and make short and long term prediction at high accuracy.

For future work, we are planning to re ne the model and introduce distributed learning of big data mining from multiple houses in a near real-time manner.

REFERENCES

[1] P. Groves, B. Kayyali, D. Knott, and S. V.Kuiken, “The big data’revolution in healthcare:Accelerating value and innovation,” 2016.

[2] M. Chen, S. Mao, and Y. Liu, “Big data: A survey,” Mobile Networks and Applications, vol. 19, no. 2, pp. 171–209, 2014.

[3] P. B. Jensen, L. J. Jensen, and S. Brunak“Mining electronic health records: towards better research applications and clinical care,” Nature Reviews Genetics, vol. 13, no. 6, pp. 395–405, 2012.

[4] D. Tian, J. Zhou, Y. Wang, Y. Lu, H. Xia, and Z. Yi, “A dynamic and self-adaptive network selection method for multimode
communications in heterogeneous vehicular telematics,” IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 6, pp. 3033–3049, 2015.

[5] M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, C. Youn, "Wearable 2.0: Enable Human-Cloud Integration in Next Generation Healthcare System," IEEE Communications, Vol. 55, No. 1, pp. 54–61, Jan. 2017.
[6] M. Chen, Y. Ma, J. Song, C. Lai, B. Hu, "Smart Clothing: Connecting Human with Clouds and Big Data for Sustainable Health Monitoring," ACM/Springer Mobile Networks and Applications, Vol. 21, No. 5, pp. 825C845, 2016.
[7] M.Chen, P. Zhou, G. Fortino, "Emotion Communication System," IEEE Access,DOI:0.1109/ACCESS.2016.2641480, 2016.
[8] M. Qiu and E. H.-M. Sha, "Cost minimization while satisfying hard/soft timing constraints for heterogeneous embedded systems," ACM Transactions on Design Automation of Electronic Systems (TODAES), vol. 14, no. 2, p. 25, 2009.
[9] J. Wang, M. Qiu, and B. Guo, "Enabling real-time information service on telehealth system over cloud-based big data platform," Journal of Systems Architecture, vol. 72, pp. 69– 79, 2017.
[10] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health using analytics to identify and manage high-risk and high-cost patients," Health Affairs, vol. 33, no. 7, pp. 1123–1131, 2014.
[11] L. Qiu, K. Gai, and M. Qiu, "Optimal big data sharing approach for tele-health in cloud Computing," in Smart Cloud (SmartCloud), IEEE International Conference on. IEEE, 2016, pp. 184–189.
[12] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "Health cps: Healthcare cyber-physical system assisted by cloud and big data," IEEE Systems Journal, 2015.
[13] K. Lin, J. Luo, L. Hu, M. S. Hossain, and A. Ghoneim, "Localization based on social big data analysis in the vehicular networks," IEEE Transactions on Industrial Informatics, 2016.
[14] K. Lin, M. Chen, J. Deng, M. M. Hassan, and G. Fortino, "Enhanced fingerprinting and trajectory prediction for iot localization in smart buildings," IEEE Transactions on Automation Science and Engineering, vol. 13, no. 3, pp. 1294–1307, 2016.
[15] D. Oliver, F. Daly, F. C. Martin, and M. E. McMurdo, "Risk factors and risk assessment tools for falls in hospital in-patients: a systematic review," Age and ageing, vol. 33, no. 2, pp. 122–130, 2004.
[16] S. Marcoon, A. M. Chang, B. Lee, R. Salhi, and J. E. Hollander, "Heart score to further risk Stratify patients with low time scores," Critical Pathways in cardiology, vol. 12, no. 1, pp. 1–5, 2013.
[17] S. Bandyopadhyay, J. Wolfson, D. M. Vock, G. Vazquez-Benitez, G. Adomavicius, M. Elidrisi, P. E. Johnson, and P. J. O'Connor, "Data mining for censored time-to-event data: a bayesian network model for predicting cardiovascular risk from electronic health record data," Data Mining and Knowledge Discovery, vol. 29, no. 4, pp. 1033–1069, 2015.
[18] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction," Data Mining and Knowledge Discovery, vol. 29, no. 4, pp. 1070–1093, 2015.
[19] A. Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger, and J. V. Guttag,"Incorporating temporal ehr data in predictive models for risk stratification renal functionDeterioration," Journal of biomedical informatics, vol. 53, pp. 220–228, 2015.
[20] J. Wan, S. Tang, D. Li, S. Wang, C. Liu, H. Abbas and A. Vasilakos, "A Manufacturing Big Data Solution for Active Preventive Maintenance", IEEE Transactions on Industrial Informatics, DOI: 10.1109/TI- I.2017.2670505, 2017.