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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 1, January 2020

Deep Learning Approach for Customer Records Prediction for Future Products Issues

Swathi Rathi

EAS SAP Analytics, Cognizant Technologies, Bangalore, India

ABSTRACT: The world of information starts with data we transform day by day. And we don't know the amount of information we pass daily may create a problem to store and utilize for further enhancements or to avoid the duplication of the information. Then big data came into existence to understand the information retrieval system and the methods of transforming information into a better way of understanding. Here we are implementing a methodology which is result oriented which can calculate the data size of data we are transferring. The size we are producing consists of the duplication of the same kind of information. We are considering cloud computing for the data storage and we are utilizing big data concepts to identify the data which was duplicated. This requires a pre-defined architecture which holds the data before publishing to the internet. This requires a theory and implementation of machine learning. Machine learning is the kind of technology which helps to predict something based on the past experiences. Here the same concept has to be used to identify the previous occurrence of the same information over the internet. This concept is used to predict the information produced by the user is genuine or not and what can be used and maintained in the repository and which can be discarded from usage from the internet. This can also be useful for identifying which kind of data is harmful for the different generation of people.

KEYWORDS: Machine Learning, Neural Networks, Tensor flow, Keras, Virtual environment, hidden layers

I. INTRODUCTION

Big data is having five kinds of V's in real time. They are Velocity which deals with the fastness of the data creation, variety deals with the which kind of data we are creating and distributing or using, volume deals with the size of information we are transferring or getting after the mining. Variability deals with inconsistency of the dataset we considered and in how many ways we can alter the dataset we considered. The last one is the veracity which deal with the trust worthiness of the information we gathered. The information we store represent the form of three kinds of data. One is structures which will be in the form of rows and columns which deals existing of the structured and organized data. We can easily understand the information we transfer and we update in the repository for the structured kind of information. This kind of structured information can gather the accurate accuracy of the predictions. This article deals with the concept of how to handle the information from the duplication. As we know data can be in many forms and the next type of data is unstructured data which consists of multimedia information. This kind of information makes the researcher challenge to maintain and predict the information which is being used for the simple things. The prediction is used for indentifying the things which are unknown and the events we perform in the prediction will lead to the future expectation of the thing. In this scenario we have to predict the information which was being published is available in the source previously. This can be an organization independent. This concept can help the people in the organization which can restrict the users to re-publish the repeated content which already existed in the repository. This concept requires few things with respect to machine learning, big data analytics and cloud computing.

The requirements of the concept were mentioned as below with the clear explanation. The lateral part of the article deals with the explanation of different things related to the prediction and identifying the information which was not required maintain in our repositories further. In prior we have to understand the technologies which are focusing on some prediction mechanisms with respect to the data gathering and managing the data in perfect manner. For an instance consider a prediction model using classification problem. In this classification problem we have to predict the new mail which was in inbox is spam or ham. There are several instances we need to consider and we have to classify

DOI: 10.15680/IJIRCCE.2020. 0801007

that mail and predict it spam or ham. If it is spam then there is no point of maintaining that mail in our repository and we can remove that from the list. In the same way we need to classify the data we have in our repository based on two key points. To publish or not and also to maintain or not. There are few further instances we need to take care and they are as follow in the next sections.

The lateral part of the section deals with background work which explains the literature survey, next explains the importance of machine learning implementation in big data applications, explaining new classification architecture to classify among the data, we support our discussion with some sort of results and finally we conclude.

II. LITERATURE REVIEW

Literature review will discuss about the background and previous work on this kind of works in the real world scenario and the information we gather from different repositories will be taken from other secondary resources over the internet. Internet will consists of different things related to each and every part of information. Starting with how to remove nails to how to launch a rocket. There is redundant information in internet which is causing the effect of gathering and maintaining the large amount of information in different repositories. Different publishers discuss in a different way on a same concept but the ideology will be same. Some of the publishers will copy the same content of others and use for their page in internet. This kind of redundant information is being created and also creating a problem in the search criteria of anything. For an instance consider we have to search for anything related to the word “Paris”. There are related search results which will be more than 1 million. Then we have to identify the required result of our search. But if we try to observe have the duplicate information on one search keyword which explains the same kind of information in all formats. For example we need to search an image related to Paris then we have tons of images which represents the result but the thing here is we are considering. But the common thing in that result is that it is speaking about only one thing. The information it produced is the redundant. The search result can be optimized and it can be used without the redundant information.

For an instance we can consider another rule for the same purpose. If we are having any person starting with the name Paris it will show the same person as the search results which cause some confusion to the user in understanding the search result. This creates a problem in maintaining some information related to the common things. In real time scenario we need to consider the our own perspective achieve any task. Here we use machine learning and big data analytics to perform predicting and storage and data manipulation.

Some of the researchers are following cloud storage for the data processing and storage. For an instance consider that we have leveraged and any kind of cloud storage and we need to give the access roles to the people who are required to get access and who are eligible to get access for the content. We have different approaches of making data clusters and making them to understand themselves to act in certain manner when they have a triggered problem. This is an interesting task to identify which among the data available is not worthy to maintain and before publishing indentifying the information which is already existed in the internet without any extra modifications and updates.

III. PROPOSED ARCHITECTURE

This proposed system consists of few things to get understand the present requirement of the system. The present requirement of the system is to understand the importance of the data which we are including in the repository based on the priority of the information and we need to slice he information and we need to gather the additions of the information with the previous information. The main motto here is as follows:

i. Need to check whether the data which we are uploading is already existed in the repository

ii. If the data is not available need to create the category of the information and present them to the repository with appropriate category

iii. If the data related to the same category is already available in the repository then we need to check for whether any additional information is added or not

iv. If the additional data is not available in the existing data then we need to add the information to the existing data and update in the same repository.

v. If the existing data and new data is not matching more than 30% then it could better to update the new information irrespective of existing data on same category

The following are technologies which we are using for this kind of implementation for handing the duplicate and redundant data in the repository.

i. Machine Learning

Machine learning is the concept of indentifying the future of a problem based on the past experiences and the present conditions. In this article we are considering natural language processing which is a subpart of machine learning implementations.

a. NLTK

NLTK is natural language toolkit is the machine learning repository for understanding the human generated information and make a decision on the input the human gave. If we consider a generala conversation between human and computer like ELIZA NLTK is used to analyse the conversation between those two users. The below image will be helpful to understand the NLTK library

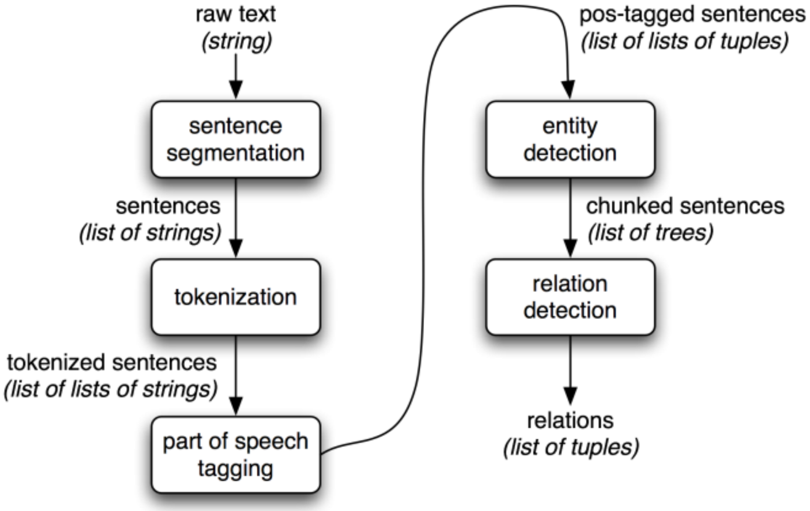

Figure 1: Explains the NLTK tokenization flowchart

b. Decision Trees

Decision trees are the basic machine learning algorithm which generates the rules based on the human decisions. The decisions we make will be based on the human instant emotions and we have to process the information based on the specific rule based process and in this scenario we have to split the information into slices and we have to identify the process of spreading the key words in the article which is being uploaded in the internet. Figure 2 indicates the sample example of decision tree algorithm.

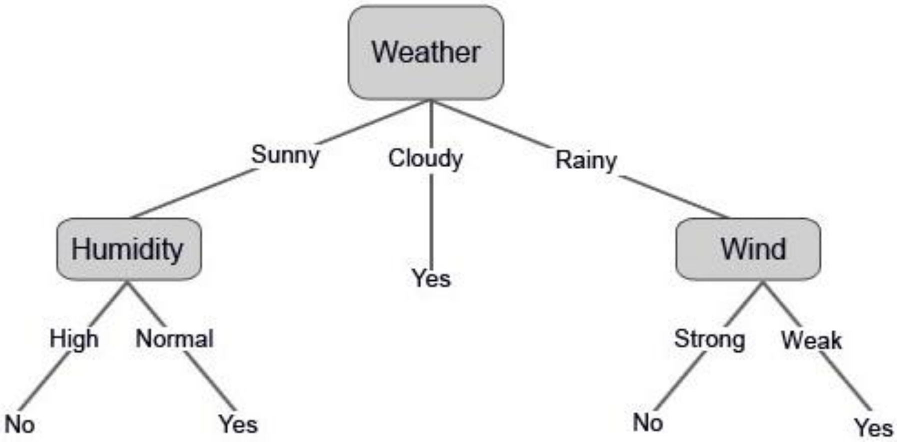

Figure 2: Decision Tree sample implementation for whether prediction

c. Map Reduce

Map Reduce is a two way task in the data manipulation and in which we perform the Map operation to split the data into chunks and done some key value pair mapping with the input data we gathered. These key value pairs are furthered used for the classifying the information related to the input dataset. Figure 3 explains the architecture of mapreduce methodology

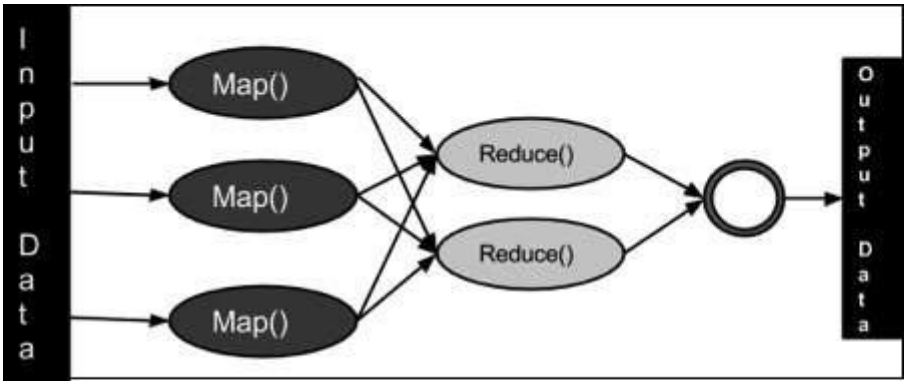

Figure 3: Map Reduce methodology

In this methodology each key value pair is further divided into the chunks based on the rules we generated to gather the information. This kind of information is furthered used to reduce the complexity and reverse mapping the information which is already present in the repository

d. AWS Cloud

AWS cloud provides default services for the big data analytics and we perform the information retrieval and storage in the cloud platform to gather and update every piece of information which is relevant to the concept. This kind of architectures help the people to free access of all the major services. The following image 4 represents AWS cloud platform for the proposing concept

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 1, January 2020

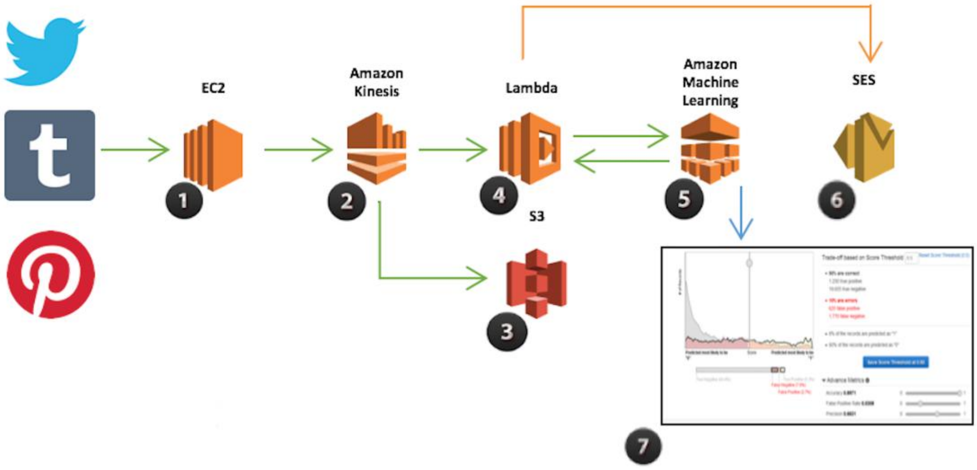

Figure 4: AWS services for big data analytics

IV. PROPOSED ARCHITECTURAL IMPLEMENTATION

This kind of proposed architecture consists of some of the things to be considered before implementing. The rules of this concept is to make the data in an understandable format and using of unwanted symbols and images can make a mess in data and this kind of noisy information have to removed before performing any kind of machine learning and big data analytics approach. This kind of approach is to make the data management some what simple and implementable. In this kind there are different phases which we are considering for better implementation of this kind of architecture. This kind of new architectures need some space of implementation whenever we are updating the information to the repository. This kind of breakage can help to build path for the solution.

In this phases first phase as follows: The first phase is to identify the category of the data which we are uploading. The category is defined based on the inputs we are gathering from the user input. The input defines some of the key words which can further used for decryption of the category of the information which we are uploading. This kind of small information can be helpful identifying the further information in a large scale. This scale of information is further stored in the cloud repository. This cloud repository is access secured. This kind of access security can be helpful for identifying the users who are using the information and who are accessing the database for any data manipulations and further tasks.

The second phase is to identify the map reduce function which can calculate the events which are happened with key value pair. This kind of key value pair is used to identify the which kind of information is being transferred over the internet regularly.

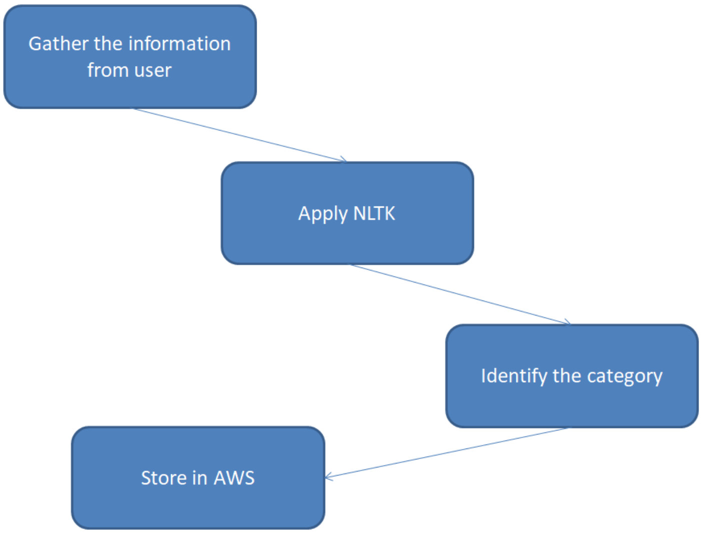

Figure 5: Phase 1

Third is to map the currently unloadable information with the previously available information. This kind of previously available information is tracked based on the keywords which the information is speaking about. We need to map those kind of key value pairs and transmit the information. Figure 5, figure 6 and figure 7 defines the architectures.

The next phase is explained in the image as follows

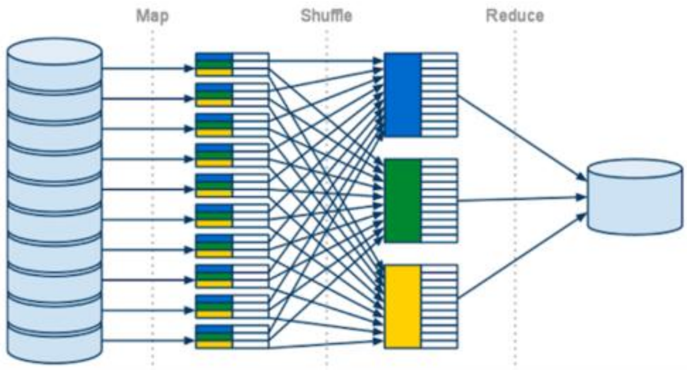

Figure 6: Map Reduce Key Value Pair

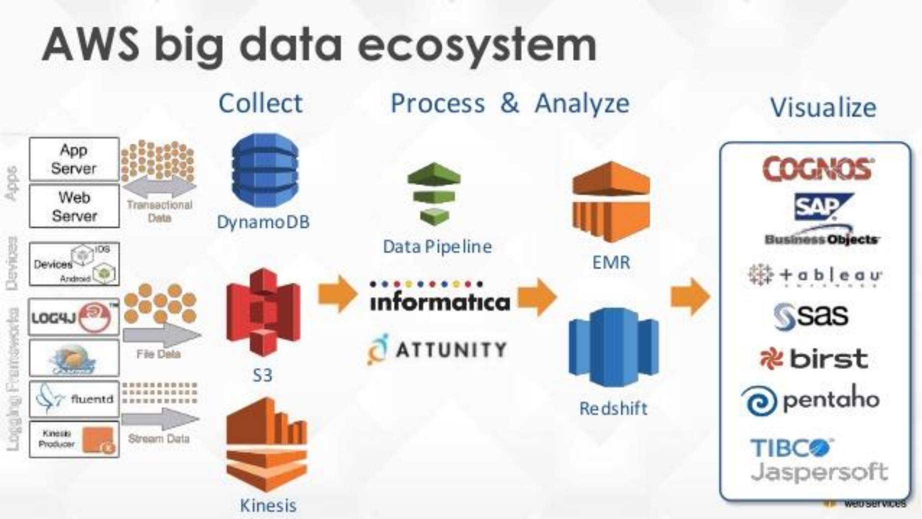

Figure 7: AWS architecture for Big Data

These are the three phases in which we are required to identify the task of updating information which was redundant in the repository.

V. RESULTS

The algorithm which we considered worked well with in collaboration with big data and the cloud computing. Random forest performed well in accuracy matters and the previous work of this kind of architectures is based on the things which are not even required to be considered like implementing the naive baysian network. This kind of networks can handle the information but the NLTK implementation will be done with Random Forest. This kind of small information can be useful for prediction of some of the accuracies of implanting the above concept with other machine learning algorithms. The table as follows in table 1. The concept hereto indentifies the words which are already used in the combination of the sentences. In this scenario we have to choose the concept which can be further used for the implementation of the analysis of the words which are being repeated in the document or the content and this can identified and checked whether anywhere in the internet we have the same combinations. This combination process will be done by the NLTK. In NLTK we implemented the slicing of the content and these contents can be helpful for identifying the small things which are majorly used in the text classification. Here we are classifying the text whether the combination is working or not. We choose different combinations of the text words and check with the global repository which was working on the same format. This format should be unique in order to accept for the uploading of the content. This kind of contents can be helpful to understand the importance of managing the redundant information. This kind of redundant information is no way used and the maintenance of the data will be much expensive. In order to manage this kind of bugs in the real time we need to manage the content with the sample requirements of the data analysis. The result we get here is a confusion matrix with the following output so that we can get the idea of implementation of the content while uploading over the internet.

Figure 8 explains the confusion matrix for the number of word combinations repeated before rewriting the content and after that.

This kind of images can be useful for implementing the content based on the word combinations and sentences.

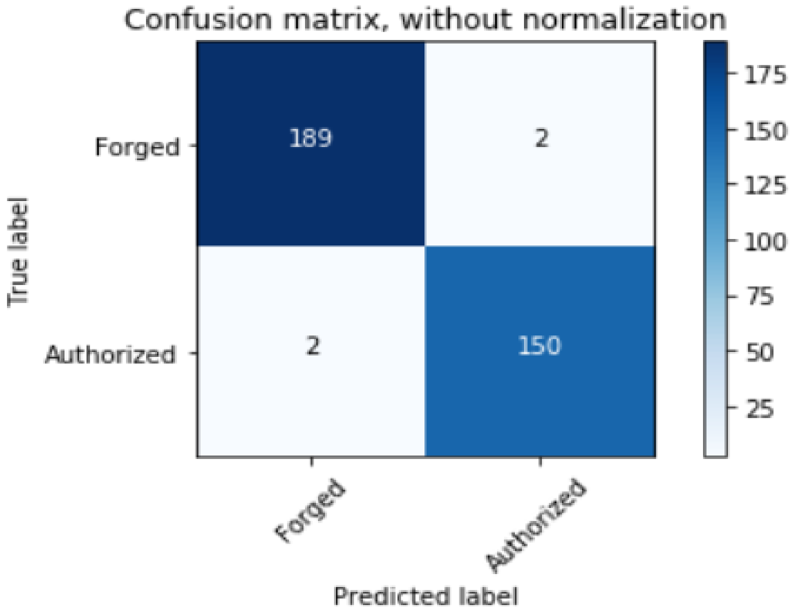

Figure 8: Confusion Matrix for the word count without normalization. This count indicates the combinations we found.

Here authorized indicates the combination of words which are not copied and the redundant and the forged indicates the words which are copied from other sources and cannot be republished over the internet.

Table 1: Accuracy table

S. No	Algorithm	Accuracy
1	Decision Trees	92%
2	Random Forest	97%
3	CNB	85%

VI. CONCLUSION

The information we store in the repository may contain the duplicate information which can be calculated and identified and eliminated. This article main focus is on such kind of things which are focusing on identifying the duplicate information in single repository which is causing some of the issues with handling the data. Then here comes to the picture of machine learning and AWS with collaboration with big data analytics. This kind of analytics performs to identify which kind of information can be saved and stored in the repository and which kid of information can be eliminated. Random forest is the algorithm achieved highest accuracy and the CNB classifier having lowest among those.

REFERENCES

[1] Kumar, A., & SAIRAM, T. (2018). Machine Learning Approach for User Accounts Identification with Unwanted Information and data. International Journal ofMachine Learning and Networked Collaborative Engineering, 2(03), 119-127.

[2 Rawat K., Kumar A., Gautam A.K. (2014) Lower Bound on Naïve Bayes Classifier Accuracy in Case of Noisy Data. In: Babu B. et al. (eds) Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012. Advances in Intelligent Systems and Computing, vol 236. Springer, New Delhi DOI: https://10.1007/978-81-322-1602-5_68.

[3] Abhishek.et.al (2020) A novel hybrid approach of SVM combined with NLP and probabilistic neural network for email phishing International Journal of Electrical and Computer Engineering 10(1):44600 DOI: 10.11591/ijece.v10i1.pp486-493

[4] C. M. Wollschlager and A. R. Conrad, ``Common complications in critically ill patients,'' *Disease-a-Month*, vol. 34, no. 5, pp. 225_293, 1988.

[5] S. V. Desai, T. J. Law, and D. M. Needham, ``Long-term complications of critical care,'' *Critical Care Med.*, vol. 39, no. 2, pp. 371_379, 2011.

[6] N. A. Halpern, S. M. Pastores, J. M. Oropello, and V. Kvetan, ``Critical care medicine in the United States: Addressing the intensivist shortage and image of the specialty,'' *Critical Care Med.*, vol. 41, no. 12, pp. 2754_2761, 2013.

[7] Gopinadh Sasubilli,Uday Shankar Sekhar, Ms.Surbhi Sharma, Ms.Swati Sharma, "A Contemplating approach for Hive and Map reduce for efficient Big Data Implementation" 2018 Proceedings of the First International Conference on Information Technology and Knowledge Management pp. 131–135 DOI: 10.15439/2018KM20

[8] O. Badawi *et al.*, ``Making big data useful for health care: A summary of the inaugural MIT critical data conference,'' *JMIR Med. Informat.*, vol. 2, no. 2, p. e22, 2014.

[9] C. K. Reddy and C. C. Aggarwal, *Healthcare Data Analytics*, vol. 36. Boca Raton, FL, USA: CRC Press, 2015.

[10] Kumar. Attangudi P. Perichappan, S. Sasubilli and A. Z. Khurshudyan, "Approximate analytical solution to non-linear Young-Laplace equation with an infinite boundary condition," 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, 2018, pp. 1-5.
doi: 10.1109/ICOMET.2018.8346349

[11] A. Perer and J. Sun, ``Matrix_ow: Temporal network visual analytics to track symptom evolution during disease progression,'' in *Proc. AMIA Annu. Symp.*, 2012, pp. 716_725.

[12] Y. Mao,W. Chen, Y. Chen, C. Lu, M. Kollef, and T. Bailey, ``An integrated data mining approach to real-time clinical monitoring and deterioration warning,'' in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. 2012, pp. 1140_1148.

[13] J. Wiens, E. Horvitz, and J. V. Guttag, ``Patient risk strati_cation for hospital-associated C. Diff as a time-series classi_cation task,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 467_475.

[14] S. Saria, D. Koller, and A. Penn, ``Learning individual and population level traits from clinical temporal data,'' in *Neural Inf. Process. Syst. (NIPS), Predictive Models Personalized Med. Workshop*, 2010.

[15] R. Dürichen, M. A. F. Pimentel, L. Clifton, A. Schweikard, and D. A. Clifton, ``Multitask Gaussian processes for multivariate physiological time-series analysis,'' *IEEE Trans. Biomed. Eng.*, vol. 62, no. 1,pp. 314_322, Jan. 2015.

[16] M. Ghassemi *et al.*, ``Amultivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data,'' in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 446_453.

[17] I. Batal, H. Valizadegan, G. F. Cooper, and M. Hauskrecht, ``A pattern mining approach for classifying multivariate temporal data,'' in *Proc. IEEE Int. Conf. Bioinformatics Biomed. (BIBM)*, 2011, pp. 358_365.

[18] T. A. Lasko, ``Ef_cient inference of Gaussian-process-modulated renewal processes with application to medical event data,'' in *Proc. Uncertainty Artif. Intell.*, 2014, p. 469_476.

[19] K. L. C. Barajas and R. Akella, ``Dynamically modeling patient's health state from electronic medical records: A time series approach,'' in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015,pp. 69_78.

[20] X. Wang, D. Sontag, and F. Wang, ``Unsupervised learning of disease progression models,'' in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 85_94.

[21] M. J. Cohen, A. D. Grossman, D. Morabito, M. M. Knudson, A. J. Butte, and G. T. Manley, ``Identi_cation of complex metabolic states in critically injured patients using bioinformatic cluster analysis,'' *Critical Care*, vol. 14, no. 1, p. 1, 2010.

[22] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, ``Modeling disease progression via fused sparse group lasso,'' in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 1095_1103.

[23] E. Choi, N. Du, R. Chen, L. Song, and J. Sun, ``Constructing disease network and temporal progression model via context-sensitive hawkes process,'' in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2015, pp. 721_726.

[24] R. Pivovarov, A. J. Perotte, E. Grave, J. Angiolillo, C. H. Wiggins, and N. Elhadad, ``Learning probabilistic phenotypes from heterogeneous HER data,'' *J. Biomed. Informat.*, vol. 58, pp. 156_165, Dec. 2015.