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. 7, Issue 5, May 2019

Attribute Evaluation of Dataset Using Chi-Square Test in Rapid Miner Studio

Swathi Rathi

EAS SAP Analytics, Cognizant Technologies, Bangalore, India

ABSTRACT: Attribute evaluation is a very popular and mandatory process in data analysis. Today the data is increasing very fast. The amount of data is increasing by multiplying itself. As increased data, the analysis of data is very tough work now. And to find the appropriate prediction is very chasing. So instead of analysing whole dataset, if we could find such variables or attributes which plays major role for predicting the final outcome, then it would be very helpful in analysis. There are various such techniques for this kind of operations. In our study we are implementing one of these techniques names as Chi-Square Test, which examines the dataset and find the most appropriate attribute for finding the prediction. The attribute having maximum weight, would be the main predicting attribute.

KEYWORDS: Statistical data, Chi-square, dataset, variable, attribute prediction.

I. INTRODUCTION

A 'Household Survey' is that the manner toward gathering and work info to modify US to understand the final circumstance and express attributes of individual home or all households within the public. throughout a home survey, field scientists analysis and record realities, perceptions and encounters from the instance households that area unit illustrative of all households within the examination territory. Instruments used for gathering info incorporate a progression of inquiries, perception agendas and records of talks.
A Chi-Square Examination is a genuine assumption Examination in like manner formed as X2 Examination. It is achieved when invalid assumption is legitimate. We played out this Examination for getting finding out about the most Variable and their weight[1-4]. In our data set, to evaluate the Variables among them, we played out this kind of Examination. As outcome it re-establishes a couple of burdens nearby the Variable. From this outcome, we can envision what Variable could without much of a stretch contrast with others to anticipate the outcome.
The variable identification plays a crucial role in the accuracy of the models we use with machine learning and deep learning. In the scenario we are considering in the further section we are explaining the other scenarios which are already done with the best variable identification and the success rate of those based on our requirement. The next section will be explaining the literature reviews related to the Chi-Square mechanism and also related to other suitable mechanisms in identifying the most impeccable features which are focusing on highest accuracy in the outcome. The later sections will cover literature review, proposed work and some other important outcomes of the models[5-8].

II. LITERATURE REVIEW

Some of the recent findings of the accuracy measurement of the features and the models state that we have pre-defined methods for understanding the basic features which are most used in the implementation of the identification of best features.

Figure 1: Structure of the performance in the classification

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. 7, Issue 5, May 2019

The above image states the important of the features extraction and its process. We have the group of features and components which are with the noisy information. We need to select the features with the following methodologies:

i. Backward Elimination

Backward elimination will be considered to be the best practice of identifying the best features to increase the accuracy. The P and SL values define the requirement of identifying the best feature and we can check the accuracy of the models and the results will explain the concept of features identification[9-12].

Below we consider one sample example of information and we are trying to analyse the information of the features and what are the features which are most useful the prediction model design. Here we have five features and the last salary feature is the predicted value and it is a dependent variable.

Based on the information we acquired we tried to implement Simple linear regression and multiple linear regression. The results are as follows and which can be identified as the huge difference in the predicted value. The result is as mentioned below.

The middle image is actual dataset values and the right side we can see that there is a difference between the simple linear regression and multiple linear regression models by comparing with actual data provided with the dataset[13-15].

The process of identifying the best feature with backward elimination consists of identifying the significance level, then later fit the model with all the possible independent variables, the highest p- value features will be identified, if the p-value is more than the significance value then it can be eliminated, then repeat the same procedure with the variables which are not eliminated.

Here we are considering significant variable and p-value as the statistical terms and here just our libraries we are using[16-17].

At this point y_pred contains predicted salaries of X_test matrix. values of y_pred are already compared with actual salaries in above screenshot.

Now, as we know in multiple linear regression,

$y = b_0+b_1X_1+b_2X_2+b_3X_3+\dots+b_nX_n$

we can also represent it as

$y = b_0X_0+b_1X_1+b_2X_2+b_3X_3+\dots+b_nX_n$ where $X_0 = 1$

So we can add one column with all values as 1 to represent b_0X_0.

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. 7, Issue 5, May 2019

Department	WorkedHours	Certification	YearsExperience	Salary
Development	2300	0	1.1	39343
Testing	2100	1	1.3	46205
Development	2104	2	1.5	37731
UX	1200	1	2	43525
Testing	1254	2	2.2	39891
UX	1236	1	2.9	56642
Development	1452	2	3	60150
Testing	1789	1	3.2	54445
UX	1645	1	3.2	64445
UX	1258	0	3.7	57189
Testing	1478	3	3.9	63218
Development	1257	2	4	55794
Development	1596	1	4	56957
Testing	1256	2	4.1	57081
UX	1489	3	4.5	61111
Development	1236	3	4.9	67938
Testing	2311	2	5.1	66029

Figure 2: Dataset considered as sample

```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       2.146e+04   4689.376      4.576      0.000    1.18e+04    3.11e+04
x1         -1421.4255   2683.353     -0.530      0.601   -6959.595    4116.744
x2            92.8656   2735.524      0.034      0.973   -5552.978    5738.710
x3             3.3361      2.410      1.384      0.179      -1.637       8.310
x4          -423.7607   1282.271     -0.330      0.744   -3070.237    2222.716
x5          9437.2530    510.978     18.469      0.000    8382.645    1.05e+04
==============================================================================
```

First Elimination

Figure 3: Sample output for Backward Elimination

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. 7, Issue 5, May 2019

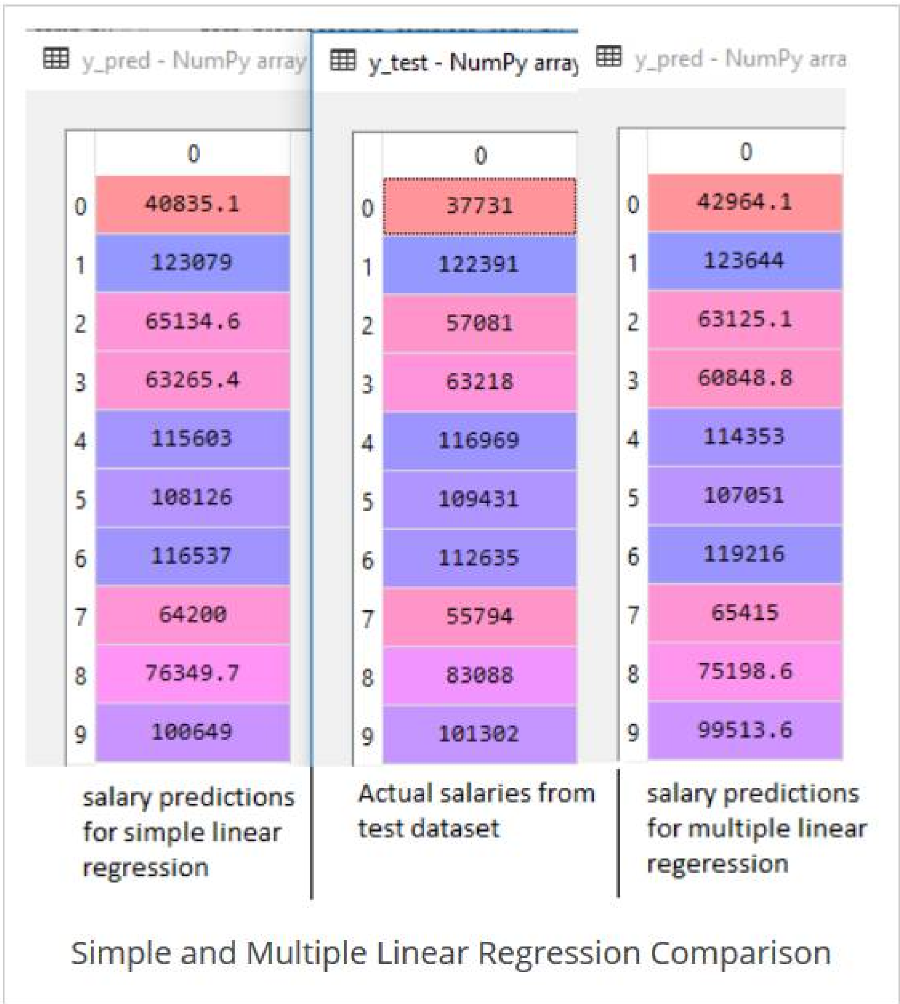

Figure 4: Comparison of the Linear Regression Models

The above image will give the implementation of backward elimination.

ii. Stepwise Regression

In statistics, stepwise regression includes regression models in which the choice of predictive variables is carried out by an automatic procedure.

Stepwise methods have the same ideas as best subset selection but they look at a more restrictive set of models.

Between backward and forward stepwise selection, there's just one fundamental difference, which is whether you're starting with a model:

- With no predictors (forward)
- With all the predictors. (backward)

At each step:

- we're not looking at every single possible model in the universe that contains k predictors such as in best subset selection but we're just looking at the models that contain the k minus 1 predictors the we already chose in the previous step.

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. 7, Issue 5, May 2019

- we're just going to choose the variable that gives the biggest improvement to the model we just had a moment earlier.

iii. Overfitting and Underfitting

Backward and forward eliminations cannot guarantee for the best output of the result with selecting the features with the ratio of P and Sl will not give the complete result of implementation. This cannot be achieved single handily. So we are into implementation of overfitting and underfitting. Whenever we are in the need to features we shouldn't add them, that comes under overfitting. When we didn't reach the required things that comes under underfitting [18].

$$\begin{aligned} ModelSpace &\approx 1 + p + (p-1) + (p-2) + \cdots + (p-k) \approx p^2 \\ &= 1 + \frac{p(p+1)}{2} \end{aligned}$$

The above computation theory can help the implementation of both stepwise implementation of backward and forward elimination proves with over fitting and under fitting of the features.

III. PROPOSED WORK

In our study we'll analyze the dataset with two main phases. One phase is to demonstrate the data using visuals in which data will be visualized in graph format which consist of some important statistical measures such as data type, max value, Min Value, Average, deviation and missing values if present. This would be helpful to identify the data whether it is complete or contains errors or incomplete data. And another work will be done in our study is that we will apply the chi-Square Test on the dataset variables to find the predicting attribute having maximum weight. The Chi-Square technique for attribute evaluation is one of the most significant techniques for attribute evaluation. In previous studies the prediction was gathered through evaluating each and every attribute of the dataset. But in our study we find only the major attributes those performs best for predicting the outcome[19-20].

(i) **Database Review using Graphics**: Here we are displaying summarized capture for each Variable with its stats.

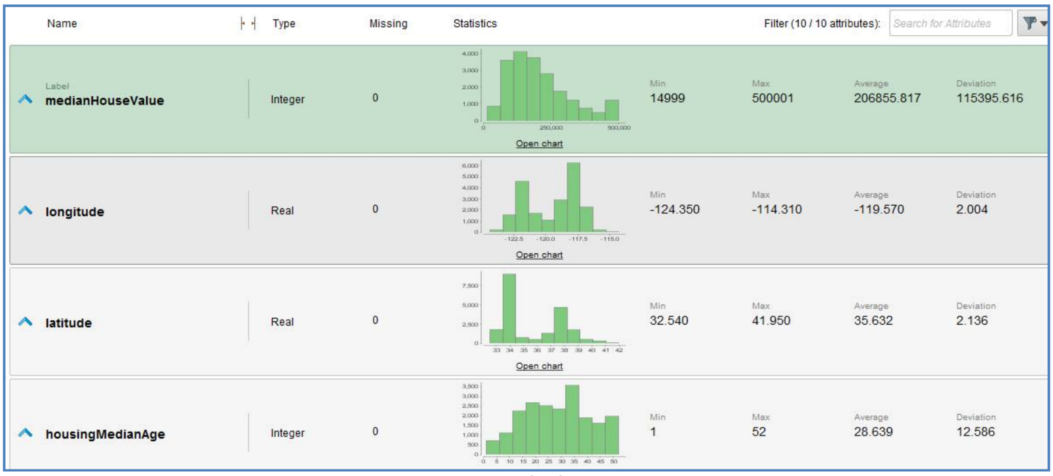

Figure 5: brief of Variables

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. 7, Issue 5, May 2019

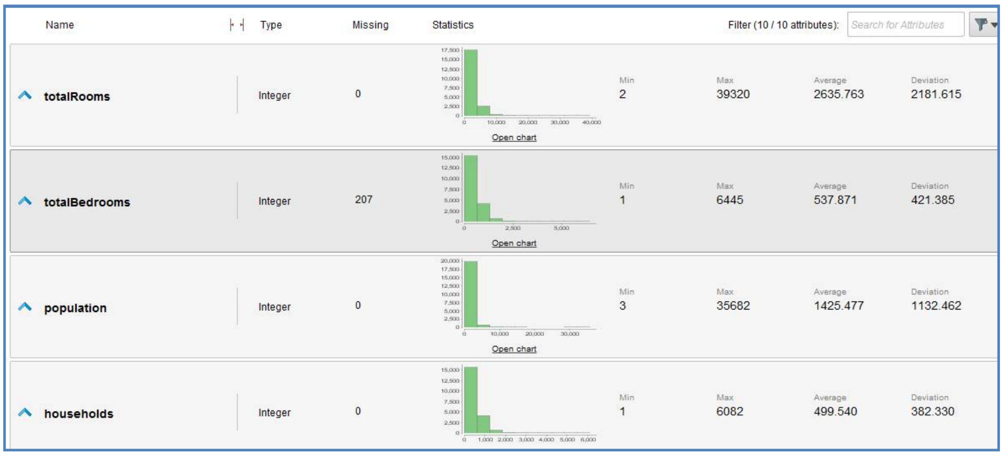

Figure 6: Summary of Variables

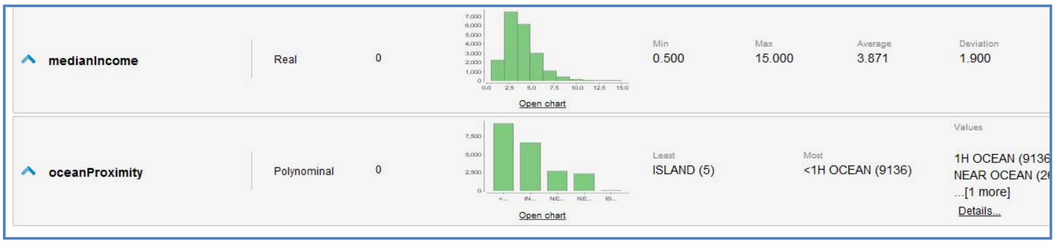

Figure 7: Summary of Variables

Table 2.1: Basic Outcomes of EDA

Attribute Name	Maximum	Minimum	Average	Std. Deviation	Mode	Missing Values	Invalid values
medianHouseValue	500001.0	14999.0	206855.816	115395.616	0	0	0
longitude	-114.31	14999.0	-119.569	2.004	0	0	0
latitude	41.95	32.54	35.631	2.136	0	0	0
housingMedianAge	52.0	1.0	28.639	12.585	0	0	0
totalRooms	39320.0	2.0	2635.763	2181.615	0	0	0
totalBedrooms	6445.0	1.0	537.870	421.385	0	207	0
population	35682.0	3.0	1425.476	1132.462	0	0	0
households	6082.0	1.0	499.539	382.329	0	0	0
medianIncome	15.0001	0.499	3.870	1.899	0	0	0
OceanProximity	<1H OCEAN (9136)	ISLAND(5)	-	-	-	0	-

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. 7, Issue 5, May 2019

The raised table portrays the Basic outcomes of the EDA methodology, which involve huge quantifiable data of each Variable. The quantifiable data joins the Variable name, Maximum estimation of Variable in educational file, Minimum estimation of Variable in instructive gathering, Average estimation of Variable in enlightening file, Std.Dev. of the estimation of Variable, Mod, Lost Values of Variable in educational accumulation and Invalid estimations of Variable in educational gathering.

In our outcome, we will have better perception about the information. In raised table we've examined that the Variable full scale Bedrooms have most outrageous lost characteristics for instance 207. We name Median house estimation as outcome ant Variable, which we have to envision dependent on various Variables of instructive file. he illuminating social event combines the Spatial information named as “longitude”, degree of house as geo spatial -area. With the assistance of result we can expect that the maximum house coordinated on the longitude of - 114.31 and degree of 41.95.Most of the houses are set where sea district is <1H and Minimum house are set close to ISLAND. When we take a gander at the database we can't foresee this result. We looked Maximum individuals of the houses is 35682.0 and least is 3.0. this conveys the database of house has Maximum masses is 35682 in house.

Chi-Square Examination: Here we’ll examine the database variables for its best predictability by implementing the Chi-Square Test technique. We will discretize the dataset variables and takes the discredited variable to the Chi Square operator.

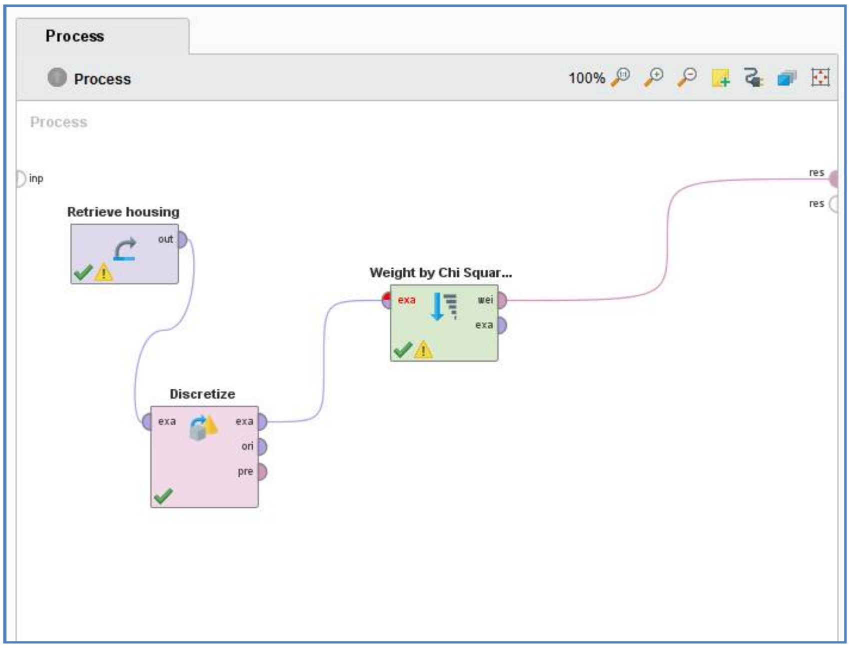

Figure: 8: Simulation Process for Chi-Square Examination

In the raised figure we've demonstrated the technique dummy to lay out the Result of Chi-Square Examination. It joins the whole data sets as data, Chi-Square Technique and Disparage Technique. Since the Chi Square Examination is achieved on simply statistical data, anyway we have a polemical Variable, we used disparage Technique [21]

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. 7, Issue 5, May 2019

attribute	weight
medianIncome	6239.950
oceanProximity	4826.509
longitude	2624.393
latitude	2412.217
housingMedianAge	292.099
totalRooms	208.570
households	125.687
totalBedrooms	97.305
population	8.388

Figure 9: The Outcome of Chi-Square Examination

The raised table demonstrates the Variable name and their heaps. As ought to be evident that the Variable median income has most noteworthy burden than various Variables, which exhibits that this Variable is most critical for envisioning the house estimation. As this thought we can sort or find any 4 to 5 dynamically huge Variables from the whole database. It is known as the Variable appraisal.

IV. RELATED WORK

The main purpose of this article is to make the researchers or the students who are working for the information retrieval and make predictions over it can understand the importance of the feature selection. For the instance consider any data like text and image. For text we can perform backward elimination and till now its in the use. Now we are proposing the new concept and explained the output we achieved in the previous section and happy for the results we achieved. In this scenario we can consider image for a change. If we are considering the image, we cannot make it as a labelled feature. So we are considering image pooling. Image pooling is the scenario of calculating the pixel values from the image and identifying which are not required in the modelling and removing them from the model. For example consider the below figure 10 as example [22].

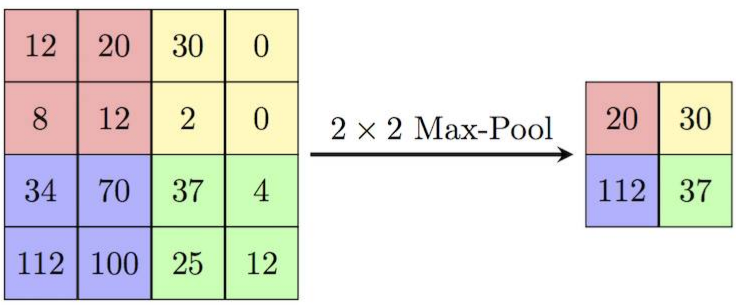

Figure 10: Image pooling with matrix.

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. 7, Issue 5, May 2019

In this scenario mentioned in the above image we can consider the pixels which are worth to be considered and maintain the matrix with the sum and average of the instances of the images.

V. CONCLUSION

The relationship between each and every feature define the strength of the model in the prediction scenario. The ratio between the features and the importance of the modelling helped to design this work and we are happy to present you the feature analysis using the chi- square analysis and the predictions are performed using the P and SL values in the backward elimination process and the process of implementation was explained clearly in the proposed work section of this article. We conclude that chi-square analysis is the best one to be followed and all the symbols and the information provided to the model in the form of features are furthered check with the importance.

REFERENCES

[1] K. B. To and L. M. Napolitano, ``Common complications in the critically ill patient,'' *Surgical Clinics North Amer.*, vol. 92, no. 6, pp. 1519_1557, 2012.
[2] C. M. Wollschlager and A. R. Conrad, ``Common complications in critically ill patients,'' *Disease-a-Month*, vol. 34, no. 5, pp. 225_293, 1988.
[3] 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.
[4] 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.
[5] A. E. W. Johnson, M. M. Ghassemi, S. Nemati, K. E. Niehaus, D. A. Clifton, and G. D. Clifford, ``Machine learning and decision support in critical care,'' *Proc. IEEE*, vol. 104, no. 2, pp. 444_466, Feb. 2016.
[6] 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.
[7] C. K. Reddy and C. C. Aggarwal, *Healthcare Data Analytics*, vol. 36. Boca Raton, FL, USA: CRC Press, 2015.
[8] D. Gotz, H. Stavropoulos, J. Sun, and F. Wang, ``ICDA: A platform for intelligent care delivery analytics,'' in *Proc. AMIA Annu. Symp.*, 2012, pp. 264_273.
[9] 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.
[10] 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.
[11] 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.
[12] 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.
[13] 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.
[14] 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.
[15] 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.
[16] 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.
[17] 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.
[18] 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.
[19] 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.
[20] 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.
[21] 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.
[22] 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.