JAY TRIVEDI

Student ID: 218449725

Enterprise Business Intelligence

Assignment 2

TIme series forecasting for hospital data.

Contents

[Abstract 2](#_Toc50991661)

[Introduction to the data analytic application background, motivation and aim 2](#_Toc50991662)

[Summary of the data set. 2](#_Toc50991663)

[Data mining technique used 3](#_Toc50991664)

[Advantages and disadvantages 5](#_Toc50991665)

[Performance evaluation and demonstration 6](#_Toc50991666)

[Conclusion 12](#_Toc50991667)

[References 13](#_Toc50991668)

# Abstract

Machine learning comprises of vast data mining algorithms that are being used in various sectors of the industries to get to the point results and predict the future values. This paper comprises of different data mining algorithms that are being used in the field of machine learning which provides accurate results depending upon vivid datasets that are being used in it. Secondly, one of the important aspects that this paper covers is the advantages and disadvantages of different data mining algorithms to get a better understanding of which algorithm suits better to the respected dataset. Lastly, an evaluation will be imposed on the dataset to get the performance evaluation results upon different attributes and algorithms and a conclusion will be drawn based on it.

# Introduction to the data analytic application background, motivation and aim

As multiple data mining algorithms are being used in different fields of the industry in the world, this research paper is focused on Time-series data analysis in the healthcare sector. One of the benefits of deploying this algorithm in the field of healthcare is that it easily identifies the nature of the event represented and accurately predicts the future forecast. Because of this, the healthcare industry will be equipped with its functionalities for the future events, for example, this COVID-19 pandemic that the world is facing today, when proper data are being collected and forecasted for this type of event then all the hospitals can improvise their systems and clinical involvement that will prove beneficial for the affected patients and future enhancements as well. The aim and motivation behind this paper are to deploy different data mining algorithms on two hospital datasets and observe which one is more accurate to another. By getting the accuracy of the respected algorithm it’ll provide a better understanding of time series forecasting which will eventually prove beneficial for the healthcare sector along with different industry fields. At the end with the help of a comparison process, it will be made clear which algorithm works better on the respected data set.

# Summary of the data set.

In this report, two datasets have been used

1. Horton General Hospital.
2. Diabetic data

**Horton General Hospital data**

This data consists of seven attributes which are index, cardio, respo, hypo, adm, year and month. All the data attributes are of a numeric value and have a total size of 4 kb.

**Data pre-processing**

After the opening of the WEKA Explorer, the data is being loaded into the pre-processor by the “open file” tab.

After opening up the dataset all the attributes are seen under the “select attribute” window in the pre-process tab where the data type, missing attributes and unique attributes are shown which are the important functionalities that will be used further in data pre-processing. Now after observing the data the index attribute was of the minimal importance in the dataset hence it was being removed by using the “Remove” filter to make the dataset more compact.

Now among all these attributes, it has to be decided which attribute is the uttermost of importance to perform the time series analysis. To do so the principal component analysis is to be done.

By applying the principal component under “select attributes” component and ranker search method of the search method we got the analysis of it. By precise observation of the Correlation matrix, ranked attributes and Eigen vector-matrix their values are 0.70 for hypo and 0.65 for adm. These two attributes or features has the larger value as compared to the rest of the attributes hence they are being used for the time series forecasting.

**Diabetic data**

The dataset consists of 50 attributes in which all the attributes are of the generic data type. The total size of the file is 18,711 kb.

**Data pre-processing**

The file is loaded into the pre-processing upon which all the important functionalities such as data type, uniqueness, instances and missing values are being noted. Upon pre-processing it was being noted that the data set consists of string values and non-specific values which were being dropped with the “remove” filter. As a result, the dataset now was of 13 attributes and 4,332 kb in size.

Upon doing this the principal component analysis was being done. Upon precise observation of the correlation matrix and ranked attributes and Eigen vector-matrix two attributes were chosen i.e. number\_output and number\_input whose values were 0.85 and 0.64 respectively.

# Data mining technique used

According to BIRNBAUM (2004), the data mining technique can be classified further as supervised mining (classification) and unsupervised mining (clustering) methods. In this report, supervised mining or classification algorithms are being used. This method is used when the input attributes are being used for prediction or forecasting purposes of the target value, one of the methods used in this research paper is the time series forecasting technique which predicts the future trend of the given attribute. Whereas in unsupervised mining i.e. clustering technique the algorithms are used on the target values with no known values to predict.

**Algorithms**

Horton general hospital

1. Linear regression.
2. Sequential minimal optimization (SMOreg).

Diabetic data

1. Linear regression.
2. K- nearest neighbour.

**Three classification algorithms are being used on two different datasets**.

One of the advantages of using the prediction algorithms is that it gives the predicted target values based on the given attribute. Such prediction algorithms used are Linear regression, SMO regression algorithms and K nearest neighbour algorithm.

**Linear regression algorithm**

According to Yan and Su (2009), the importance of using the linear regression algorithm is to evaluate the impact of the predictor variable on the respected outcome value. Whereas the word “regression” is used to determine the relationship between the two variables of the dataset.

**Horton general hospital**

**Steps for linear regression**

1. At first, the dataset is being loaded into the WEKA processor upon which all the functionalities are being observed such as datatype, uniqueness and missing values.
2. Upon further all the useless attributes are drop-down in this case will be the “index” attribute which consists of the numeric values which are not useful for this algorithm.
3. Thirdly, the principal component is being deployed under the “select attributes” tab and “ranker” search method.
4. After deployment of the above the co-relation matrix, ranked attributes and eigenvector components are being observed.
5. They give insights which attribute is of the utmost importance as compared to other attributes.
6. In this hypo and adm are being used whose values are 0.70 and 0.65 respectively.
7. After assessing these attributes, the “forecast” tab is used under which the specific algorithm is being deployed. At first, the filtered classifier is used under which “linear regression” and “remove” method is opted under “classifier” and “filter”. After this, the performance evaluation is being selected upon which will give us RMSE and mean values of the opted attributes.
8. After running the above a specific value are obtained but further modifications are required to get a reduced RMSE value.
9. Now the linear regression algorithm is opted under the “base learner configuration” along with it the “use custom lags” options is being selected with “remove leading instances with lagged values” option under it.
10. After running this the RMSE value difference is huge as compared to before hence further the “evaluation on held-out training” is opted under the evaluation tab.
11. Now after running with the above changes it reduces the RMSE value gives two predictions for “future training set” and “future test set”.

**Steps for Sequential Minimal Optimization (SMO regression) algorithm**

According to Platt (1998), sequential minimal optimization algorithm opts to solves any small problem with the best optimization problem occurring at every step. Along with it also uses two Lagrange multipliers analytically. As well as the algorithm doesn’t store any extra matrix storage which gives the results faster. This algorithm is deployed against the linear regression algorithm is the SMO regression model. The above steps remain the same for the SMO regression model, the only difference that is being observed for this model is the change in the RMSE value and its predictions.

**Diabetic data**

**Linear regression algorithm**

1. The linear regression process for this data is alike what we observed above, the diabetic dataset is opened in the WEKA Explorer. With this, all the important functionalities are observed.
2. Along with is all the unnecessary attributes are dropped with the help of remove filter in the processor panel to reduce the size of the data to get an accurate forecast.
3. Thirdly, the principal component analysis is being done on the dataset upon which the correlation matrix, ranked attributes and eigenvector matrix is being observed.
4. Upon this, we’ll get two values with uttermost importance depending upon the values that are 0.85 and 0.64 for numer\_output and number\_input.
5. Keeping the respected attributes in mind the forecast tab is being used where all the steps from “step 7 to step 11” are similar from above.

**K- nearest neighbour algorithm**

According to Batista and Monard (2001), the purpose of deploying k- nearest neighbour algorithm is that it proves beneficial in predicting both the discrete attributes as well as continuous attributes. Along with it, this algorithm has the ability of easy adaption with any attribute as a class by simply modifying the attribute for the distinct metrics. This algorithm is to be used as an alternative algorithm for the linear regression in the diabetic dataset. **The steps for this algorithm are similar to above i.e. from step 1-5.**

# Advantages and disadvantages

**Linear regression algorithm** (Zanutto 2006)

Advantages

* The algorithm is easy to implement because of its interaction with other covariates as well as better interpretation with the output coefficients.
* Less complex than other algorithms.
* Gets influenced by overfitting but gives better results when the dimension of the dataset is reduced.

Disadvantages

* The model assumes the existence of linearity between the two variables that are the dependent variable and independent variable.
* Another disadvantage is that if the number of observations is less than the number of features this model should not be used as it leads to overfitting which considers noise when building the model.

**Sequential minimal optimization regression model (SMOreg)** (Pham et al. 2017)

Advantages

* Give better application results when the data is unknown
* Compatible with structured and semi-structured data such as text and images.
* Kernel function plays an important role in this model that solves all the complex problems.

Disadvantages

* Takes longer amount of training time for larger datasets.
* Opting for a better kernel function is not as easy as compared to other models.
* Difficulty in the understanding of the final model including weights and variables.

**K nearest neighbour algorithm** (Batista & Monard 2001)

Advantages

* Easy to implement and use
* Addition of new data can be done at any point
* Doesn’t need any training period

Disadvantages

* Takes longer duration of time for a larger dataset
* Difficulty in working with high dimensions of the dataset
* Feature scaling is needed at certain steps of the algorithm
* Doesn’t give better results if the data consists of missing values and outliers.

# Performance evaluation and demonstration

The performance evaluation simply means all the algorithms are being compared based on the dimensionality of the dataset such as the size of the dataset, instances and attributes etc. in this research paper a total of three algorithms have been used which are linear regression, sequential minimal optimization regression model and k nearest neighbour algorithm on two datasets. The datasets are of different sizes in which one of them is of Horton general hospital which is relatively in lesser size as compared to the diabetic data.

**Mean absolute error**

According to Willmott and Matsuura (2005), the mean absolute error is a relatively simple as it includes the summation of the absolute values of the errors obtained from the dataset and then dividing the total error with the value “n”. it forecasts the set of errors without considering their directions as well as it includes all the individual differences that are weighing equally in the average.

**Root Mean Squared Error (RMSE)**

According to Willmott and Matsuura (2005), the calculation of the RMSE is obtained with the help of three simple steps. Firstly, we get the total squared error first as a series of summation of each squared error which means each error impacts the total error in proportion to its square as compared to small errors. The large errors obtained has a larger impact on the total squared error as compared to small errors, which means as there is an increase in the large errors the total squared error will also show the increase in its impact. Secondly, the total square error is then divided by “n” which leads to mean square error (MSE). Lastly, the RMSE is taken as the square root of the mean square error (MSE). Hence the lower the value of RMSE the better fit as compared to larger RMSE values it is for the algorithm of that desired dataset.

**Horton general hospital**

**Linear regression algorithm**

Training data

A screenshot of a computer

Description automatically generated

As shown in the above figure the RMSE value for hypo in the training dataset is 0.29 whereas for adm is 56.19 this is obtained by configuring the lag creation as well as the evaluation tab.

Test data

A screenshot of a social media post

Description automatically generated

From the above figure, we can see that the RMSE value for hypo and adm are 0.20 and 53.19 in the test data where you can observe the drop in the values as compared to the training dataset.

**Forecasting**

**Linear regression model**

**Training data**

A screenshot of a social media post

Description automatically generated

The future forecast predictions on the training dataset are shown above.

**Test data**

A screenshot of a cell phone

Description automatically generated

The future predictions for the test data are shown above.

**Sequential minimal prioritization regression model (SMOreg)**

**Training data**

A screenshot of a social media post

Description automatically generated

The above RMSE value is obtained by doing configurations in the lag creation as well as in the evaluation tab where the values obtained are 0.31 for hypo and 57.7 for adm.

**Test data**

A screenshot of a social media post

Description automatically generated

According to the figure obtained above the RMSE value for test data is 0.20 and 53.5 for hypo and adm respectively which is comparatively lower than the training data.

**Forecasting for SMO regression model**

**Training data**

A screenshot of a social media post

Description automatically generated

**The future forecasting for the training data is shown in the above diagram.**

**Test data**

A screenshot of a social media post

Description automatically generated

The future forecast for the test data is shown in the above diagram

**Evaluation between both models**

The results obtained from the above two algorithms it has been noted, the difference between their RMSE values for both the training and test data can be seen. It can be precisely noted down that the RMSE values for the linear regression model give more accuracy as compared to sequential minimal optimization regression model (SMOreg) where the values are slightly higher. Hence opting for a linear regression model is better than SMOreg based on the above accuracies.

**Diabetic data set**

**Linear regression model**

**Training data**

The linear regression algorithm when applied to the diabetic data gave the RMSE of 1.25 and 1.26 for the number\_outpatient and number\_inpatient in the training data.

A screenshot of a social media post

Description automatically generated

**Test data**

The RMSE values for the number\_outpatient and number\_inpatient are 0.88 and 0.68 respectively for the test data which is a reduced version as compared to the training data.

A screenshot of a social media post

Description automatically generated

**Forecasting**

**Training data**

A screenshot of a social media post

Description automatically generated

The future forecasting for the training data is shown above.

**Test data**

A screenshot of a social media post

Description automatically generated

The future forecasting for the test data is shown above.

**K nearest neighbour algorithm**

**Training data**

When the k-nearest neighbour algorithm is applied to the diabetic dataset it was observed that the RMSE values were null for the training data.

A screenshot of a computer

Description automatically generated

**Test data**A screenshot of a social media post

Description automatically generated

From the above calculations, the RMSE values were slightly higher as compared to the training data which are 0.95 and 0.70 for number\_outpatient and number\_inpatient.

**Evaluation**

Two algorithms were being deployed on the diabetic dataset (i.e. Linear regression and K-nearest neighbour) and results were obtained which showed that k nearest neighbour has shown more accuracy for both the datasets as compared to the linear regression algorithm. The sequential minimal optimization algorithm was also deployed on the dataset but one of the drawbacks of that algorithm is that it doesn’t function properly on a larger dataset. Hence choosing the k-nearest neighbour algorithm suits best for this dataset.

**Forecasting**

**Training data**

A screenshot of a computer

Description automatically generated

The future forecasting of the training data is shown above.

**Test data**

A screenshot of a computer

Description automatically generated

The future forecast for the test data is shown above.

# Conclusion

The machine learning stream has multiple data mining algorithms already deployed which can be used on different datasets and the aim of the organisation or personnel. In this research paper, we saw the time series forecasting method being deployed in the healthcare industry i.e. Horton general hospital and the diabetic data. Two algorithms were being deployed on each of them which has shown certain accuracies depending upon the size of the dataset and the attribute used for prediction. From the results obtained it can be seen that the implementation and accuracy of the algorithms are solely based on the size of the dataset and the aim of use.

# References

Batista, G & Monard, MC 2001, 'A study of K-nearest neighbour as a model-based method to treat missing data', in *Argentine Symposium on Artificial Intelligence*.

BIRNBAUM, EBD 2004, 'Application of data mining techniques to healthcare data', *Infection control and hospital epidemiology*.

Pham, BT, Bui, DT, Prakash, I, Nguyen, LH & Dholakia, M 2017, 'A comparative study of sequential minimal optimization-based support vector machines, vote feature intervals, and logistic regression in landslide susceptibility assessment using GIS', *Environmental earth sciences*, vol. 76, no. 10, p. 371.

Platt, J 1998, 'Sequential minimal optimization: A fast algorithm for training support vector machines'.

Willmott, CJ & Matsuura, K 2005, 'Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance', *Climate research*, vol. 30, no. 1, pp. 79-82.

Yan, X & Su, X 2009, *Linear regression analysis: theory and computing*, World Scientific.

Zanutto, EL 2006, 'A comparison of propensity score and linear regression analysis of complex survey data', *Journal of data Science*, vol. 4, no. 1, pp. 67-91.