A Novel Approach for Predicting the Length of Hospital Stay With DBSCAN and Supervised Classification Algorithms

Panchami V U, N. Radhika
Amrita Vishwa Vidyapeetham
Coimbatore, India
n radhika@cb.amrita.edu, panchamivu@gmail.com

ABSTRACT: *Patient length of stay is the most commonly employed outcome measure for hospital resource consumption and to monitor the performance of the hospital. Predicting the patient's length of stay in a hospital is an important aspect for effective planning at various levels. It helps in efficient utilization of resources and facilities. So, there exist a strong demand to make accurate and robust models to predict length of stay. This paper analyzes various methods for length of stay prediction, its advantages and disadvantages and proposes a novel approach for predicting whether the length of stay of the patient is greater than one week. The approach uses DBSCAN clustering to create the training set for classification. The prediction models are compared using accuracy, precision and recall and found that using DBSCAN as a precursor to classification gives better results.*

Keywords: Classification Aalgorithms, Supervised Learning, Hospital Management, Patient Monitoring, DBSCAN, Resource Management, Training Set

Received: 18 January 2014, Revised 28 February 2014, Accepted 4 March 2014

1. Introduction

Patient Length of Stay is an important performance indicator of a hospital which provides a better perception of the resource consumption and flow of patients through a healthcare system. LOS estimations has a lot of applications in operational and clinical functions of a healthcare system such as finding out the future bed usage, making estimates of the forthcoming demands on different hospital resources, defining the case-mix, providing help to the patients to understand the course of the disease and recovery, finding health insurance schemes and reimbursement systems in the private sector, planning discharge dates for elderly patients, patients who are dependent, patients with needs and as a crucial factor for the quality of life of the patients and families.

Thus, a prediction model that can predict the length of stay of a patient can be an effective tool for the healthcare providers for making proper plans for preventive interventions, to perform better health services and to do the management of hospital resources more efficiently. With the help of the accurate estimation of the stay of patients, the hospital can plan for more efficient resource utilization. Predicting the probable discharge dates can help to estimate available bed hours, that results in higher average occupancy and less waste of resources in the hospital.

This paper proposes a novel approach for predicting whether the length of stay of a patient in a hospital is greater than one week. The approach identifies groups of similar hospital claims from the dataset using DBSCAN clustering and utilizes these

groups as the training set to classify the LOS of patients with high accuracy. Four different prediction models are created using Logistic Regression, Neural Network, Support Vector Machine and Naive Bayes and are compared using performance metrics of accuracy, precision and recall. The training approach that produces the best performance is found out. The statistical significance of this result is analysed using t test. The outcome of this study is a training approach that produces prediction models with better accuracy, precision and recall.

2. Related Work

The studies on finding out how long patients will stay in a hospital have happened extensively since 1960s by building prediction models.

The study on hospital length of stay by Gustafson [1] developed and compared five methodologies for prediction of length of stay. Three of these methodologies gave a point estimate of the length of stay based on physicians' subjective opinions, while the other two gave a probability distribution over all lengths of stay based on empirical data. The data used for the study were collected from a sample of eight inguinal herniotomy patients and were stratified over four categories of length of stay. Since the data set was too small, the study didn't give good performance.

Woods et al [2] evaluated the predicted and actual length of stay in 22 Scottish intensive care units using the APACHE III system. The performance of the APACHE III length of stay prediction was evaluated and found that it stratifies patient populations effectively but is a poor predictive tool for individual patients.

Liu et al. [3] conducted a study to predict LOS by building linear regression models based on a data set of hospitalizations from 17 hospitals in Northern California. To predict the outlier LOS, longer than a week, they used logistic regression. In this model, Laboratory Acute Physiology Score (LAPS) and Comorbidity Point Score (COPS) were added which were found to have excellent predictive value. These scores are calculated when a patient is hospitalized using different measures. The results showed that adding LAPS and COPS data to the models improved LOS predictions.

Kulinskaya et al. [4] conducted the study for two purposes. The first purpose was to find out factors that influence LOS. The factors considered were: admission method, discharge destination, provider hospital type, NHS region and speciality (acute or geriatric). The second purpose was to determine the most accurate and consistent statistical method for analysing LOS and other similar data. The study compared many linear regression-based and maximum likelihood-based models, and found truncated maximum likelihood (TML) to have the best fitness value. However, the accuracy of the predictions made by this model was not discussed in the paper.

Steven et al [5] used neural network for evaluating the level of illness of trauma patients and for predicting the length of stay accurately. The work built prediction models based on backpropagation, radial-basis-function and fuzzy ARTMAP algorithms. The data used for the study of pediatric trauma patients are collected within the first ten minutes of the arrival of that patient. Neural network performed well for this medical domain problem. For predicting patient length of stay, the best performance was achieved by backpropagation network. In evaluating the level of injury of a patient, fuzzy ARTMAP showed superior performance. The study recommended the combination of backpropagation and fuzzy ARTMAP to produce optimal combined result [5].

Ali Azari et al. [6] proposed a multi-tiered data mining approach for Predicting Hospital Length of Stay (PHLOS). The study used K-means clustering on the raw data in order to select representative training tuples. After clustering, training sets are created by taking samples close to the cluster centers. Another training set was derived from the raw data without clustering. Classification was then performed using different classifiers and different training datasets. The study found that usage of clustering as a precursor, to form the training set, gives better prediction results as compared to non-clustering based training sets. And BNet, SMO and JRip generate relatively better LOS predictions. This study used K-means clustering to derive training sets. But K-means clustering fails to address outliers. It does not perform well in situations where the cluster sizes and density differs, and also where the clusters belong to some non-globular shapes.

3. Proposed System

This study proposes a data mining approach using DBSCAN clustering technique as a precursor to classification task for

predicting whether the length of stay of a patient in a hospital is greater than one week. Doing clustering before classification helps in eliminating noise points and getting important data points from the large data set. Good training set can produce better accuracy. Since DBSCAN handles outliers and takes care of clusters of different sizes, shapes and densities, this approach proposes the use of DBSCAN clustering. The samples for training set are selected from the clusters. The test set is created in such a way that the samples in the training set and test set are non-overlapping.

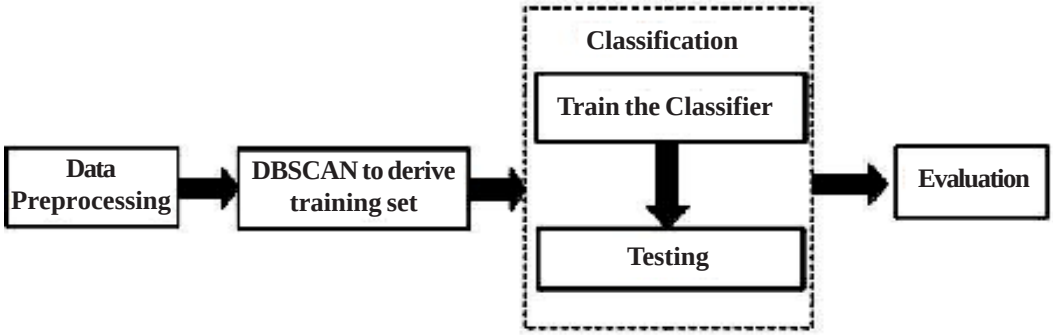

Figure 1. Architecture of the proposed system

After obtaining these sets, classification is done. The study takes four classification algorithms-Logistic Regression, Neural Network, Support Vector Machine and Naive Bayes. The performance of the classifiers are obtained by three performance measures-Accuracy, Precision and Recall. The overall architecture of the proposed system is given below.

The system consists of mainly four modules-data preprocessing, clustering, classification and evaluation. The classification module is divided into two sub modules-training and testing. The evaluation module consists of evaluation of classifiers and evaluation of training approaches.

3.1 Preprocessing

Databases are highly susceptible to noisy, missing, and inconsistent data due to their huge size and their likely origin from multiple, heterogeneous sources [7]. Low-quality mining results are produced by the use of low-quality data. In order to improve the quality of the data and, consequently, of the mining results and also to improve the efficiency and ease of the mining process, data should be preprocessed. Methods used for data preprocessing are data cleaning, data integration and transformation, and data reduction. To improve the accuracy, efficiency, and scalability of the classification or prediction process, the following preprocessing steps are applied to the data.

3.1.1 Data Cleaning

Since the data tend to be incomplete, noisy, and inconsistent, data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data [7]. This study intends to eliminate the records with missing values.

3.1.2 Data Integration and Transformation

Data mining often requires data integration which involves the merging of data from multiple data stores [7]. The data may also need to be transformed into forms that are appropriate for mining.

Normalization [7] is a technique of data transformation where the attribute data are scaled so as to fall within a small specified range, such as -1.0 to 1.0, or 0.0 to 1.0. The proposal insists to do min-max normalization technique on the attributes selected for the study. They are to be normalized into the range 0.0 to 1.0. Min-max normalization technique performs a linear transformation on the original data. This technique preserves the relationships among the original data values.

3.1.3 Data Reduction

To obtain a reduced representation of the data set, data reduction techniques can be applied. After data reduction, data set of much smaller in volume, that closely maintains the integrity of the original data is obtained [7]. It has been seen that, mining on the reduced data set is more efficient [7]. Attribute subset selection is a strategy for data reduction where irrelevant, weakly

relevant, or redundant attributes or dimensions are detected and removed [7]. The data set may contain a lot of attributes which are irrelevant for predicting the LOS. So, a set of relevant attributes are to be selected from the dataset. The study insists to select the relevant attributes by considering the opinion of an expert in the field and the factors found in the literature.

3.2 DBSCAN

DBSCAN (Density Based Spatial Clustering for Applications with Noise) [8] is a density based clustering approach which is resistant to noise. It can handle clusters of different sizes and shapes. According to density based clustering, a cluster is a dense region of points, which is separated by low-density regions [8]. It is used when the clusters are irregular or intertwined, and when noise and outliers are present. The key idea is that for each point of a cluster, the neighborhood of a given radius (Eps) has to contain at least a minimum number of points, i.e. the density in the neighborhood has to exceed some threshold. DBSCAN considers three types of points-core point, border point and noise point [9]. A point is a core point, if it has more than a specified number of points (MinPts) within Eps. A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point. A noise point is any point that is not a core point or a border point.

To find a cluster, DBSCAN starts with an arbitrary point P and retrieves all points that are density-reachable from P with respect to Eps and MinPts. If *P* is a core point, this procedure yields a cluster with respect to Eps and MinPts. If P is a border point, no points are density-reachable from P and DBSCAN visits the next point of the database [10].

3.3 Classification

Classification finds out a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown [7]. The model is derived from the training data set which consist of data objects and its class label. This paper analyzes four classifiers. They are Logistic Regression, Neural Network, Support Vector Machine and Naive Bayes.

3.4 Evaluation

The various prediction models $M = fm_1, ..., m_i g$ formed are to be analyzed based on the set of performance measures $pm_{ij} = fp_{i1}$, ..., p_{ij} g generated. For the proposed approach, *j* is considered to be equal to 3, including accuracy, precision and recall [11]. The evaluation is done by comparing these performance measures for the generated models, by taking into account the clustering method used to generate training sets. The combination of clustering and classification technique that is best suited to predict hospital length of stay is extracted from the evaluation. The best training set is also found out from the comparison of performance measures. Then *t*-test is performed to analyze the statistical significance of the result.

4. Experimental Results

This study evaluated the performance of four classifiers when trained using three training sets. Logistic Regression, Neural Network, Support Vector Machine and Naive Bayes are the classification algorithms selected. In order to find good training tuples, DBSCAN is done on the dataset. The experiment is conducted with the help of R statistical package [12].

4.1 Data Description

The dataset used for the study is a randomly generated dataset for the replication of PHLOS approach [13]. It is based on Heritage Health Prize data. The data contains attributes such as primary condition group, specialty, Charlson Index and DSFS (Days Since First Stay) for predicting hospital length of stay.

4.2 Data Preprocessing

Data integration is performed based on record identifier. Many attributes present in the dataset were irrelevant to predict LOS. So, data reduction by attribute subset selection is done based on the opinion of an expert in healthcare. The next challenge was the missing values present in the data. The specification file of the database contained the encodings used to represent the missing values in each attribute. So, those values are found out, and the corresponding records are eliminated. Then the LOS attribute is converted into two different groups - '*greater than one week*' and '*less than one week*' and encoded as 1 and 0 respectively. Then the data is transformed into the range [0, 1] by using min-max normalization method.

4.3 Creation of Training and Test sets

Training set and test set creation from the whole data is done after preprocessing. Training set 1 is formed without doing any clustering on the preprocessed data. Training set 2 is formed after performing K Means clustering on the data. The value of K is selected as 2 in order to get two classes of length of stay greater than one week and less than one week. Samples that are close to the cluster centers are selected. Then DBSCAN is done on the preprocessed data. The noise points found out by DBSCAN are eliminated and thus formed Training set 3. A test set is created from the preprocessed data in such a way that records in the test set are non overlapping to the records in the training set.

4.4 Classification

The 3 training sets formed are applied to 4 classification algorithms so that a total of 12 prediction models are generated. The four classification algorithms used were Logistic Regression, Neural Network, Support Vector Machine and Naive Bayes. The test set is applied to the prediction models and the accuracy, precision and recall of the models are calculated. The results obtained are shown in Table 1.

Table 1 shows the accuracy, precision and recall of different prediction models generated. Support Vector Machine, when trained using Training set 3, gives the highest accuracy, precision and recall.

Figure 2 indicates the accuracy of Logistic Regression, Neural Network, Support Vector Machine and Naive Bayes classifiers when trained using Training set 1, 2 and 3. It shows that clustering before classification gives better accuracy. Models trained using Training set 3 produced best accuracy results.

Figure 3 shows the precision of the classifiers. It is found that clustering before classification increases the precision of the models. Dataset formed using DBSCAN produces the best precision value.

The recall of the classifiers is represented in Figure 4. Datasets generated using clustering show better results. In such datasets, DBSCAN generated dataset produces highest recall value.

4.5 Evaluation

The study used three training sets created based on three different clustering methods. The training set 1 was taken from the data set without doing any clustering. Training set 2 was created by performing K Means clustering on the data set and training set 3 was created after doing DBSCAN. From Figure 2, Figure 3 and Figure 4, it is obtained that training set generated using DBSCAN gives better prediction results. In order to find out the statistical significance of this result, unpaired *t* test [14] is performed. The *t* test compares the means of two groups. The test is done for each performance metric by taking the performance of Training approach 3 in one group and the performance of the other approach in the next group.

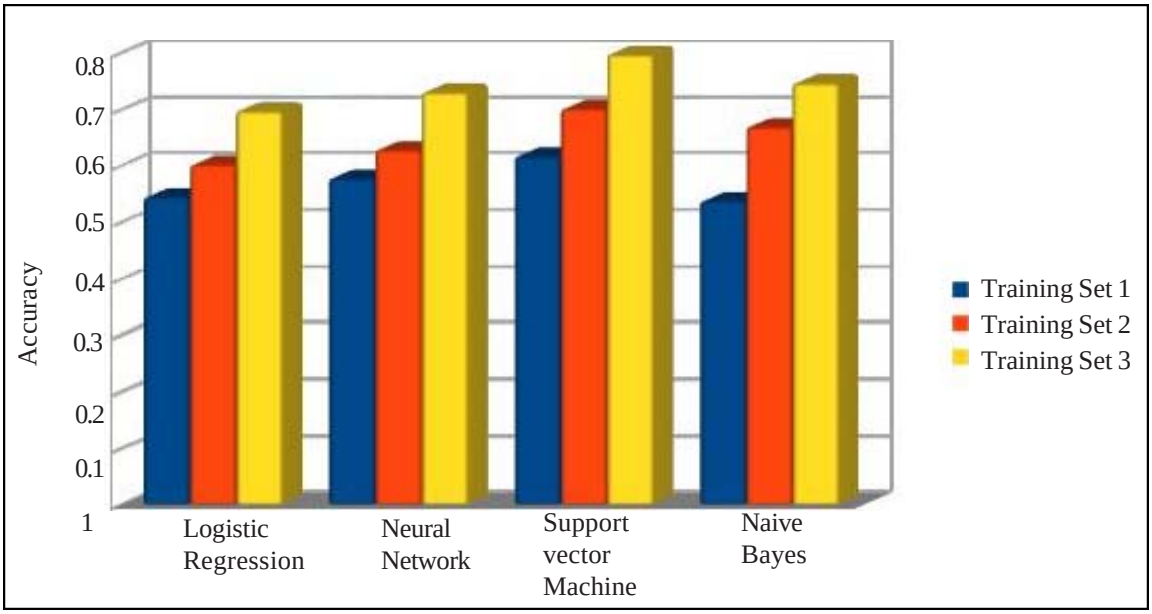

Figure 2. Graph comparing the accuracy of different classifiers

Performance metrics	Training Set 1	Training Set 2	Training Set 3
Logistic Regression			
Accuracy	0.546	0.603	0.698
Precision	0.557	0.615	0.711
Recall	0.466	0.554	0.667
Neural Network			
Accuracy	0.579	0.629	0.731
Precision	0.545	0.643	0.736
Recall	0.565	0.609	0.679
Support Vector Machine			
Accuracy	0.618	0.702	0.798
Precision	0.585	0.704	0.781
Recall	0.569	0.701	0.776
Naive Bayes			
Accuracy	0.539	0.670	0.747
Precision	0.474	0.691	0.759
Recall	0.543	0.697	0.743

Table 1. Performance Analysis of Prediction Models

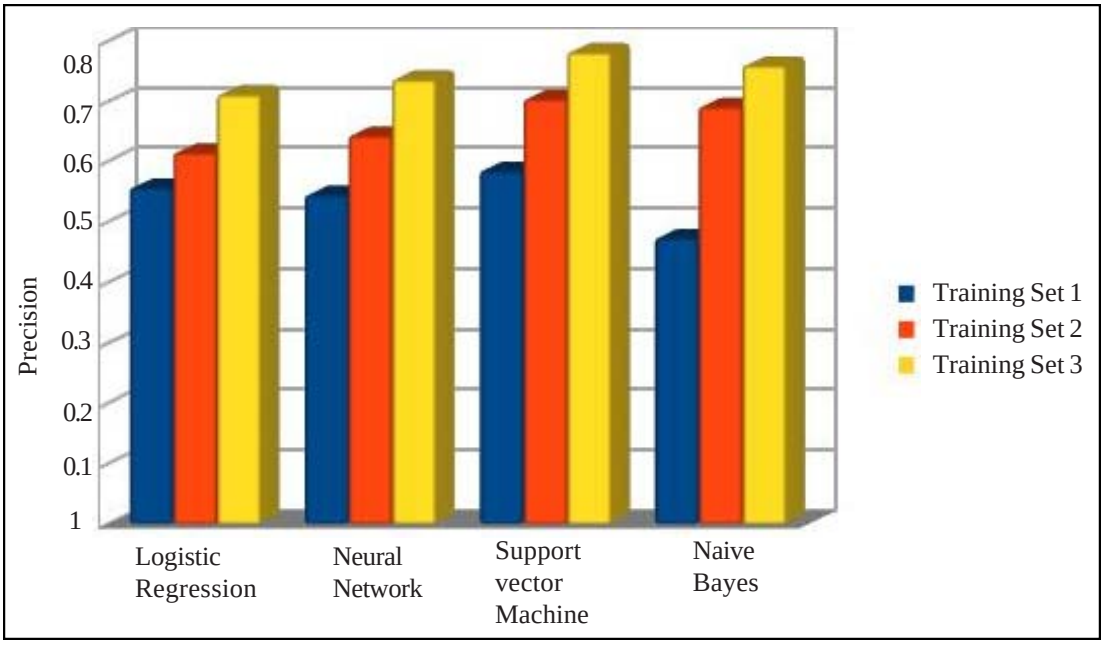

Figure 3. Graph comparing the precision of different classifiers

The first test is done with the Null Hypothesis: there is no significant difference between the accuracies of predictive models when trained using training set 1 and training set 3. The two tailed unpaired *t* test for 95% confidence interval resulted in a *p* value of 0.0033. So, the result is found to be statistically significant and the Null Hypothesis is rejected.

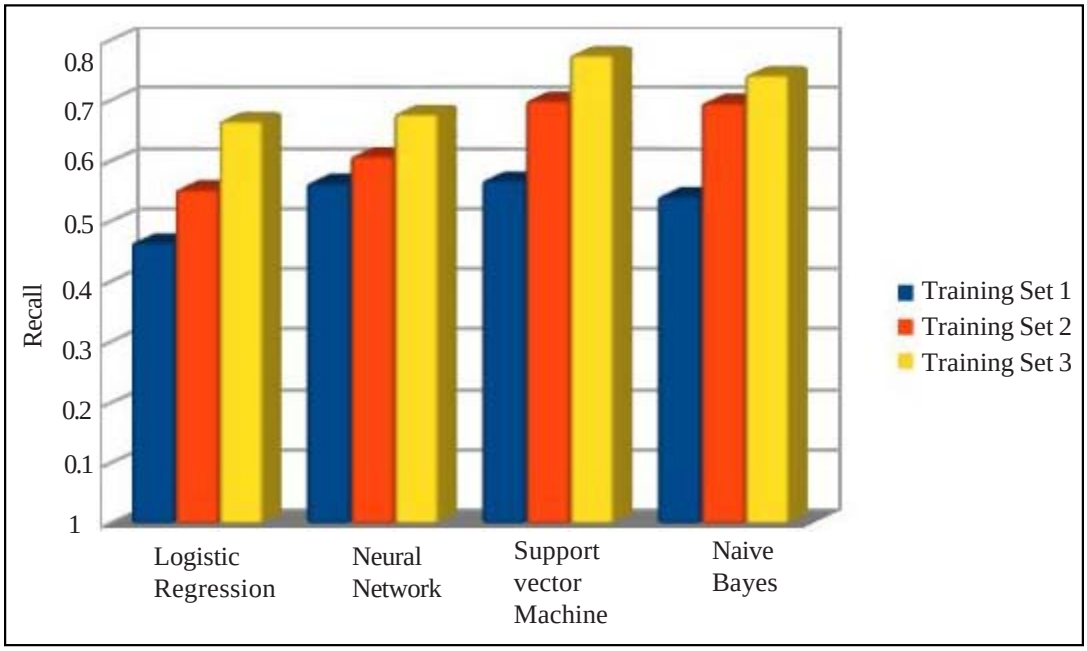

Figure 4. Graph comparing the recall of different classifiers

The next test is done by taking Null Hypothesis as: there is no significant difference between the accuracies of predictive models when trained using training set 2 and training set 3. The p value obtained after the test was 0.0509. Eventhough training set 3 outperforms training set 2, the result is not quite statistically significant. So, the Null Hypothesis is accepted.

From the two tests based on accuracy, it is observed that predictive models trained using training set 3 outperforms the predictive models trained using training set 1 significantly. But with prediction models trained using training set 2, the outperformance of predictive models trained using training set 3 is not quite significant.

In order to find out the significance of the result in terms of precision, two tests are performed. The first test took Null Hypothesis as: there is no significant difference between the precisions of predictive models when trained using training set 1 and training set 3. The test resulted in a p value of 0.0016 and the result was very statistically significant. So, the Null Hypothesis is rejected.

Then Null Hypothesis is taken as: there is no significant difference between the precisions of predictive models when trained using training set 2 and training set 3. The p value obtained after the test was 0.0453. The result is found to be statistically significant and the Null Hypothesis is rejected.

The two tests on precision show that, the outperformance of prediction models when trained using training approach 3 is statistically significant when compared with the prediction models trained using training set 1 and training set 2.

In order to find out the statistical significance of the outperformance of training approach 3 based on the performance metric recall, two tests are performed. The first test took Null Hypothesis as: there is no significant difference between the recalls of predictive models when trained using training set 1 and training set 3. The t test for confidence interval 95% resulted in a p value of 0.0082. The result is found to be very statistically significant and the Null Hypothesis is rejected.

Then the test is performed by taking Null Hypothesis as: there is no significant difference between the recalls of predictive models when trained using training set 2 and training set 3. The p value obtained was 0.0865. The result is not quite statistically significant and the Null Hypothesis is accepted.

It is observed that the outperformance of prediction models when trained using training approach 3 is statistically significant when compared with models generated by training set 1.

The *t* test performed for finding out the statistical significance of the outperformance of prediction models trained using training set 3 shows that these models are better models for predicting the hospital length of stay.

5. Conclusion

Determining how long a patient will stay in a hospital is important in healthcare to provide better care for the patient and thus to increase the reputation of the hospital. This paper focuses on length of stay prediction. Different prediction models are generated that predict whether the length of stay of a patient is greater than a week or not. These models are compared using accuracy, precision and recall. The problem is viewed as a classification task in which each patient record is classified into one of the two classes-'*greater than one week*' and '*less than one week*'. The work proposed an approach of doing clustering before classification in order to find representative training tuples and thus to increase the performance of prediction models.

A density based clustering approach called DBSCAN is used for clustering. Since it is a density based clustering, clusters of different sizes, shapes and densities are found out. It also eliminated noise points from the data set. Different prediction models are built based on Logistic Regression, Neural Network, Support Vector Machine and Naive Bayes. From the experiments conducted, it was observed that prediction models based on Support Vector Machine gives the best performance. It was also found from the results that training the prediction models using the training set created by DBSCAN approach provides the best performance.

References

[1] Gustafson, D. H. (2002). Length of stay prediction and explanation, *Health Services Research*, 37 (3) 631-645.

[2] Woods, A.W., MacKirdy, F. N., Livingston, B. M., Norrie, J., Howie, J. C. (2000). Evaluation of predicted and actual length of stay in 22 Scottish intensive care units using the apache III system, Anaesthesia, 55 (11) 1058-1065.

[3] Liu, V., Kipnis, P., Gould, M. K., Escobar, G. J. (2010). Length of stay predictions: Improvements through the use of automated laboratory and comorbidity variables, Medical care, 48 (8) 739-744.

[4] Kulinskaya, E. K., Gao, H. D. (2005). Length of stay as a performance indicator: robust statistical methodology, *IMA Journal of Man-agement Mathematics*, 16 (4) 369-381.

[5] Steven Walczak., Ronald J. Scorpio., Walter E. Pofahl. (1998). Predicting Hospital Length of Stay with Neural Networks, *In*: Proceedings of the Eleventh International FLAIRS Conference, p. 333-337.

[6] Ali Azari, Vandana P. Janeja, Alex Mohseni. (2012). Predicting Hospital Length of Stay (PHLOS): A Multi-Tiered Data Mining Approach, in 2012 IEEE 12th International Conference on Data Mining Workshops, p.17-24.

[7] Jiawei Han, Micheline Kamber. (2006). Data Mining: Concepts and Techniques, Second Edition, Elsevier.

[8] Martin Ester, Hans-Peter Kriegel, Jorg Sander and Xiaowei Xu. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, *In*: Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon.

[9] DBSCAN. Available at http://en.wikipedia.org/wiki/DBSCAN, ac-cessed on 30th April.

[10] Clustering: DBSCAN density reachability and connectivity. (2013). Available at http://rss.acs.unt.edu/Rdoc/library/fpc/html/dbscan.html, accessed on 30th April.

[11] Precision and Recall. (2013). Available at http://en.wikipedia.org/wiki/Recall and precision, accessed on 10th March.

[12] The Comprehensive R Archive Network. (2013). Available at http://cran.r-project.org/, accessed on 1st March.