LOSH Prediction using Data Mining

Ruchi Rathor
Department of Computer Science
Dr. D Y Patil SOE
Pune, India

Pankaj Agarkar
Department of Computer Science
Dr. D Y Patil SOE
Pune, India

ABSTRACT
Only when resources and time of the hospital is managed, the productivity of the Hospital services enhances. Both time and resource consumptions are at its peak when patient is admitted to the hospital. So, they can best be managed at this time of stay. Also, managing the emergency cases as they arrive should also be taken care. These factors can be managed by estimating the future resource requirements of the hospital. The rate at which resources are consumed is to be determined. Hence, if the LOSH (Length Of Stay at the Hospital) of the patient is determined, we can easily manage the resources and emergency admissions. Hence, to derive the stay duration of the patient in the hospital is an important operation. This paper proposes a prediction model that predicts the length of stay of the patient in the hospital and a solution to handle emergency situations when doctor is unavailable.

We used basic clustering methods like DBSCAN (Density Based Spatial Clustering Application Network) and K-Apriori. In addition, we compared the execution time of the both.

General Terms
Data Processing; Data Mining;

Keywords
Data Processing; DBSCAN; Apriori; K-means Clustering; Prediction;

1. INTRODUCTION
To further enhance the productivity of the hospital, resources like food, beddings, medicines, equipment etc. should be managed properly and efficiently. During the stay at the hospital by the patient, the rate of consumption of the resources is quite high. Hence, there is a requirement to know for how many days the patient may stay at the hospital. By knowing the LOSH (Length of Stay at the Hospital) of the patients, the amount of resource consumed till date and future consumption can be determined. And accordingly, future demands can be made. This way resources exploitation can be handled.

Hospitals will also be able to provide high medical services based on resource availability. Also, future appointments with the patients or further admissions of the patients can be easily planned. All of this will increase the occupancy rate at the hospital.

One more factor that affects the hospital's productivity is the handling of an emergency situation, as it arrives. For example: burn cases, accident cases etc. It might happen that as an emergency arrives the specialist is not present or might be busy handling other cases. We propose a model in which the emergency cases can be handled if the specialist is not present at the site. For example, if the burn has arrived and the staff does not know what to do the proposed system can guide them. Hence, by handling these two factors the service quality of the hospitals can be enhanced.

In addition, to these operations we analyzed and compared the time execution between DBSCAN, k-apriori algorithms. Both the approaches were used over the dataset of 9052 elements and then compared.

2. RELATED WORK
To know the prediction of LOSH by the patient, a lot of research work being done in the past, on the respective field.

D. H. Gustafson [1] compared five prediction methodologies, i.e. statistical prediction, physician's prediction of LOS at time of admission request, physician's prediction of LOS during the patient's hospital stay and discharge prediction by nurses. Out of which physician's prediction of LOS at time of admission request and physician's prediction of LOS during patient's stay gave point values and other two i.e. statistical prediction and discharge prediction by nurses gave poor precision as the data collected was incomplete and estimation result was not efficient.

V. Liu, P. Kipnis, M. K. Gould, and G. J. Escobar [2] proposed a model based on linear regression that predicted length of stay (LOS). They used diagnosed set of data from almost 17 hospitals that have 205,177 hospitalizations. In addition, they combined Laboratory Acute Physiology Score (LAPS) and Co-morbidity point score (COPS) to the linear regression model that gave improved result.

Panchami V U [3] proposed and compared 12 predictive models each one with different clustering and classification approaches. They composed three training set, two by using k-means clustering and DBSCAN algorithms and one by without clustering. And then they used four classification algorithms i.e. Linear Regression, Neural Network, SVM and Nave Bayers for treating training set; resulting in twelve predictive models. The Length of Stay at the hospital was predicted for more than seven days. They used statistical data from the hospitals. The result of each model was compared based on performance measures (accuracy, recall and precision). Out of twelve predictive models, the model that had used DBSCAN and SVM (Support Vector Machine) gave the best prediction.

Ali Azari, Vandana P. Janeja , Alex Mohseni [4] to cut down qualm of Length of Stay at hospitals, they proposed and compared forty models that are formed by using different clustering and classification approaches. They composed three training sets by using k-means clustering with different values for k. To form first training set, k was number of conditions the patient was hospitalized. For second training set, the value of k was Charlson Index, which defines comorbidity conditions. And for third training set, ks value was Idealk i.e.

the mid-point of the SSE (Sum of Square Error) curve. With these three training sets one test set was also considered, that was formed with randomly collecting the elements that are not included in the other three training sets. Then comparison was performed based on performance measures like AUC, Kappa, Accuracy, Recall and Precision. The ranking was done using Friedman Test. This proposal couldn't handle the outliers and also poorly performed for clusters of dynamic shapes and densities. Also, because of presence of anomalous tuples prediction was not truly efficient.

E. K. Kulinskaya and H. D. Gao [5] used the health statistical data from UK NHS for 1997/98 and 1998/99 for analyzing five variable i.e. admission method, discharge destination, provider type, specialty and NHS region; these can directly affect the prediction of the LOS of the patient. They have compared the Standard Method with the Robust Method. The standard method used was General linear Models (GML) and the robust method was Truncated Maximum Likelihood (TML). Out of them TML proved to be better estimator then General Linear Models. But the accuracy of the prediction made is not compared with the actual one.

3. PROPOSED SYSTEM

There are three major users that will interact with the proposed system, they are:

3.1 The Patient

All of them will register into the system before performing any operation. The patient can either register online or at the time of visit to the hospital. For the first timer patients, they have to provide their medical history along with the current medical history then only they can request for the appointment with doctor.

3.2 The Hospital Staff Member

Any one from the hospital staff can access the system. Whatever information was provided by patient during online registration will be in a general form compared to information needed at the visit. A detailed health condition is to be provided to the staff member so that initial report can be generated. Based on the initial report, resources can be reserved, if they are available. The availability of the resources will be checked and further demands will be managed by the staff. The patient will ask for appointment which will be finalized on doctor's confirmation.

3.3 The Doctor

Being the administrator of the system, all management work will be done by the doctor. Both details of the diseases and the resources are managed. The appointment requests of the patients will be first confirmed by the doctor and then finalized. The diagnosis of the patient will be done on the basis of the current health information, the medical history and the report generated by the system.

4. BASIC FUNCTIONS

The proposed system has following basic functionalities:

4.1 Registering the Appointments

The patient will send the request for the appointment which will be forwarded to doctor by the staff member. This request will be approved only on confirmation with the doctor and notified to the patient. If the appointment for that specific period is not approved, further suggestions can be provided by the doctor and notified to the patient. If the patient finds it possible to visit at that time he will confirm it.

4.2 Resource Management

The staff members are responsible for keeping track of the resource consumption and resource release. In addition to this, they are also responsible for placing future demands of the resources. During the visits by the patient, the resources are reserved based on the health condition of the patient. Once done with the prognosis and doctor have prescribed the medicine, the reserved resources are allocated to the patient. And at the time of discharge the resources are released.

For allocation of resources, Bankers Algorithm is obeyed. Banker's algorithm allocates the resources without giving raise to deadlock by first analyzing how safe the state might become after resources are assigned. An allocation of resource, might lead to deadlock and the operation might hang this is an unsafe state; hence these type of allocation are avoided by bankers algorithm.

In the proposed work, the patients will require the resources and will have to allocate without blocking the future appointments.

4.2.1 Banker's Algorithm

P: set of patients admitted to the hospital;
M(r): max. Resources requested by a patient;
C(r): current resource allocation to the patient;
R: set of resources;
Ri: resource been requested by the patient;
A(r): currently allocated resources.
Step 1: p request for Ri where $p \subseteq P$;
Step 2: if $M(r) - \{C(r) + Ri\} \leq A(r)$;
Allocate C(r) to p;
return safe;
Else return unsafe;

The resources will be categorized as Operation Theater and the Rooms as shown in fig 1. As the System predict the disease, the resources need to be reserved; this consequences into checking the availability of the required resources. The availability can be checked by either searching by the name or by the category. Also, availability of basic resources can be checked directly as the direct links are provided at the bottom of the web-page.

4.3 Prediction of length of stay at the hospital by the patient

The prediction is done based on the current symptoms and the medical history of the patient, which is provided by the patient at the time of visit. For prediction of LOSH of the patient, basic data mining approaches are used. The medical data goes through preprocessing, which comprises of three steps: data cleaning, data integration and transformation and finally data reduction. Then, with this processed data, symptoms for a disease are grouped together and based on which length of stay of the patient is predicted. This process is clustering and classification.

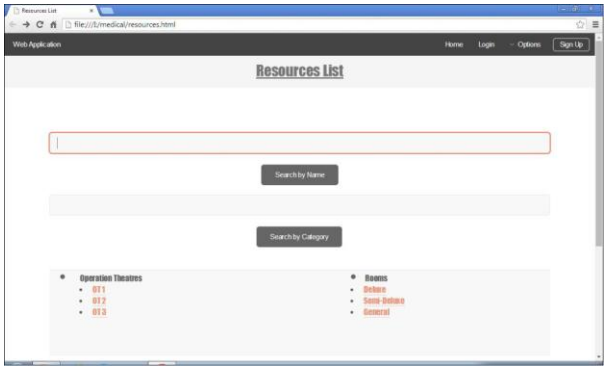

Fig 1: Check Resource Availability

The algorithms used for clustering are:

4.3.1 DBSCAN:

DBSCAN stands for Density Based Spatial Clustering for Applications with Noise[8,10]. As the name itself tells that it is a density based clustering approach which is not affected by noise. It can easily treat multiple clusters of multiple size and shape, where a cluster is a densely populated region of points that are separated by the low density regions [6]. The neighbourhood can be traced by a given radius called Eps. DBSCAN forms clusters by considering three types of points - core point, border point and noise point[8].

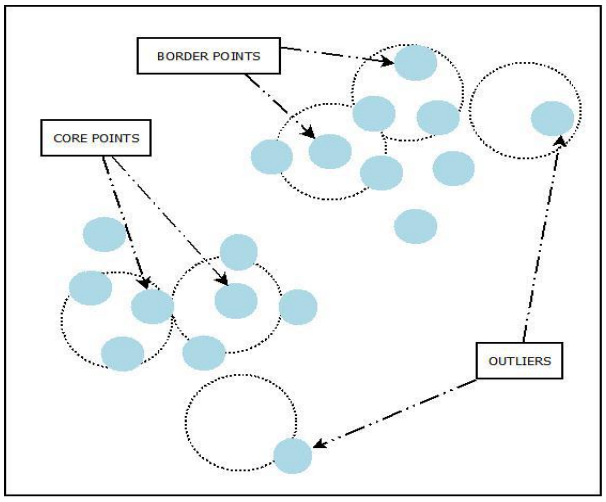

Fig 2: Points for DBSCAN

Input: P set of points in a plain;
Output: irregular clusters of points;
Step 1: select a point p;
Step 2: Retrieve all points density-reachable from p wrt € and MinPts.
If p = core point, a cluster is formed.
If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database.
Step 3: Continue the process until all of the points have been processed.
A point is a core point if it has more than a specified number of points (MinPts) within Eps. These are points that are at the interior of a cluster. A border point has fewer than MinPts within Eps, but is in the neighbourhood of a core point. If any two core points are within a distance Eps of one another then they are put in the same cluster. This way noises are eliminated. A noise point is any point that is not a core point or a border point. Any border point that is close enough to a core point is put in the same cluster as the core point. Noise points are discarded. This way noises are eliminated [6,8].Points within an Eps from a point is the epsilon-neighbourhood. Core points is a set of points under Eps-Neighbourhood contains at least MinPts of points. A point x is said to be directly density reachable, if it is directly density-reachable from point p if x is within the Eps-Neighbourhood of p and p is a core object.

4.3.2 k-apriori:

K-Apriori is the combination of Apriori and K-means Algorithm. They work as follows:

Ck → Candidate itemset of size k;
Lk → frequent itemset of size k;
L1 = frequent items;
Input: S set of Symptoms;
Output: clusters;
Step 1: for(k= 1; Lk!=null; k++) do begin
Step 2: C(k+1) = candidates generated from Lk;
Step 3: for each transaction t in database do increment the count of all candidates in
C(k+1) that are contained in t
L(k+1)= candidates in C(k+1)min support
end
return U Lk

4.3.3 K-Means Clustering:

K-Means is a Partitioning methods that freely moves the elements from one cluster to another cluster, initiating from a first initial partitioning. the number of clusters to be formed will be predefined by the user. This method, partitions the dataset into K clusters i.e c1,c2,....,ck. Each cluster is represented by its centers or means, which is calculated by taking the mean of elements that belong to respective cluster:

$\mu_k = 1/N \sum X_q$ where q = 1…..N_k

Input : S set of Symptoms;
Output : clusters;
Step 1: initialize k cluster centers;
Step 2 : while termination condition is not satisfied do
Step 3: Assign instances to the closest cluster center
Step 4: Update cluster centers based on the assignment
end while
The algorithm starts by taking an initial center of set of clusters. In each iteration, each element is given to its nearest cluster center by considering the Euclidean distance between the two. Then the cluster centers are re-calculated. The predictions will be made at least three times or if required can increase, based on patient's condition: First, at the registration level- when patient register with their details; Second at the diagnosis level when the patient visits the doctor and Third after admitting the patient. The prediction will be done again and again if the condition doesn't change.

4.4 Handle the Emergency Situation

The initial prediction will have suggestions of at least three possible diseases that the patient might suffer from. This prediction is made based on the current medical information (symptoms) and the medical history provided to system at the time of visit by the patient. The patient will be treated, as per the suggestions. For example: if an emergency case of accident has arrived and doctor is not present, then to handle this situation all the primary steps like steps to stop bleeding or pain; will be performed to relief the patient. All these initial steps will be instructed by the LOSH and followed as said; and will be easily performed by the staff.

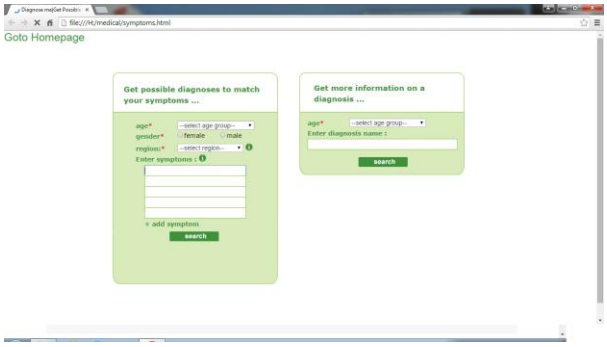

Fig 3: Systems Diagnosis Page

The patient or staff member can explore the condition in two ways: either by feeding the list of symptoms along with age group the patient belongs to or by directly feeding the disease name as shown in fig 3.

5. SYSTEM ARCHITECTURE

As we have already mentioned the number of users and there their interactions with the system, the overall structure of system comprises of four modules:

5.1 Interaction Modules:

The users will interact with the system. The basic operations will be updating the databases and accessing the data. The patient will fill in the required details into the database of the system whereas the staff member and the doctors can access this data for diagnosis and can update the results and future demands through interaction module.

5.2 Convertor Module:

The convertor module will perform data pre-processing. Databases might consist of noisy, missing, and inconsistent data as they have a very huge size and gathered from numerous heterogeneous sources. Hence, this low level quality of data along with low mining method will not give a precise result [7]. Hence, we need to improve the data quality. To improve the preciseness of the result and to ease the mining process, preprocessing of the data is to done.

The basic steps followed while data pre-processing are:

5.2.1 Data Cleaning:

Data might be noisy, incomplete and also inconsistent; these can be treated as impurities of data. These impurities are removed by a cleaning process, which fill in the missing values, determine the outliers and remove the noise and make the data consistent [7].

5.2.2 Data Integration and Transformation:

Data needed for the prediction are collected from various sources hence; data integration is the process of merging of data from different sources. As data collected from different sources will be vast, they are required to be scaled together in order to minimize the calculation complexity. So, data transformation will scale them so that they similar data will fall under same range [7].

5.2.3 Data Reduction:

Mining the reduced data gives an efficient result. Data are reduced in such a way that it maintains the integrity of the original data. For this the attribute subset selection strategy is used that detects the irrelevant, weakly relevant and redundant attributes, and remove them [7].

5.3 Data Mining Modules:

data mining module has basic implementation of clustering and classification. The k-means will cluster the common symptoms of the different diseases i.e from various cases whereas the apriori will specify the disease related to the symptoms and give the disease name. This combination of k-means and apriori is the K-Apriori.

5.4 Evaluator Modules:

The outcome of the data mining module will be evaluated in evaluator module and will be verified by the employees.

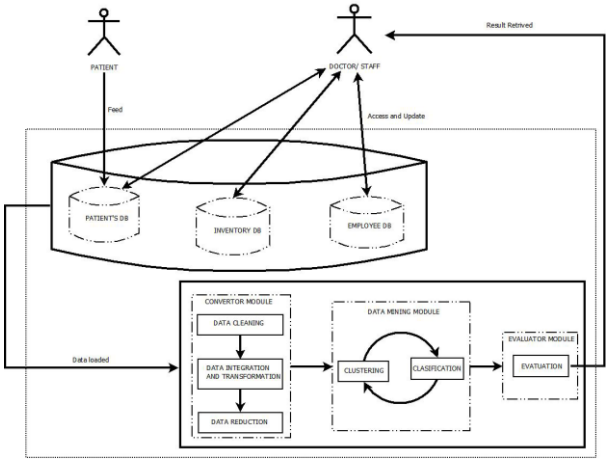

Fig 3: Systems Diagnosis Page

6. MATHEMATICAL MODEL

The proposed system is defined by a set of four tuples, as follows:

S = {Doc,Comp, Pat,Db}

where;

System, S = {LOSH} ;

Employee, E = {Doc, Comp};

Doctor, Doc = {D1,D2,D3….Dn};

Specialist, SDoc ⊆ Doc;

Compounder, Comp = {C1,C2,C3….Cn};

Patient, Pat = {P1,P2,P3…..Pn};

Database, Db = {Medb} i.e. medical database.

All the users i.e doctors, staff members (here considered only compounder) and Patients will interact with medical database of the system.

7. ANALYSIS

For the analysis of the proposed approach we analyzed and compared the execution time of DBSCAN and k-apriori algorithms. We analyzed the execution time with respect to 30 cases that has around 10,000 input values (here will be Symptoms). The input values are the set of symptoms provided by the patient. The following table shows the individual execution time of the DBSCAN and K-Apriori with respect to number of inputs (Symptoms).

Table 1. Execution Time with respect to total no of Inputs

Total no of symptoms	Apriori	K-Means	K-Apriori	DBSCAN
30	0.015	0.01	0.025	0.093
36	0.046	0.03	0.076	0.094
88	0.125	0.011	0.136	0.094
160	0.327	0.02	0.329	0.156
252	0.499	0.014	0.513	0.171
364	0.39	0.23	0.62	0.14
496	0.34	0.265	0.605	0.141
648	0.483	0.4	0.883	0.14
820	0.421	0.34	0.761	0.14
1012	0.421	0.355	0.776	0.141

8. RESULTS

We determined the execution time of K-apriori and DBSCAN independently and then compared them. The sample table shows the independent execution of Apriori, K-Means algorithm, K-apriori and DBSCAN with respect to total number of inputs(Symptoms).Both the algorithms were treated with same number of inputs i.e 10,000 and with same values. When compared, we found that execution time of DBSCAN is comparatively very short then K-apriori. But as we go on increasing the number of inputs the execution time of DBSCAN increases exponentially whereas there is no change in K-apriori.

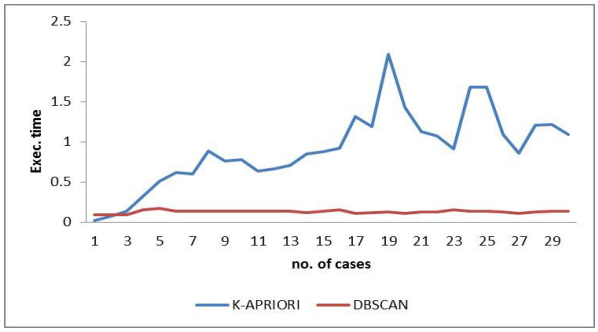

Fig 5: DBSCAN vs K-Apriori

The longest execution time for DBSCAN is 0.171 seconds whereas K-apriori has 2.087seconds as shown in fig 5.

9. CONCLUSION

As Resources of the hospital needs to be managed properly there is a need to predict the LOSH of the patient. We proposed a system in which prediction of LOSH is the prime objective along with a solution to control the emergency cases. To accomplish the prediction, we have used data mining techniques i.e DBSCAN and K-apriori algorithms for clustering. We determined and compared the execution time of DBSCAN and K-Apriori to know which one will be a better and fast. We observed that among the two, DBSCAN proved to be faster than the K-Apriori, but increases exponentially as number of inputs increases.

10. ACKNOWLEDGMENT

We would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. Also, my family and friends for constantly encouraging and supporting me.

11. REFERENCES

[1] D. H. Gustafson, Length of stay prediction and explanation, Health Services Research, vol. 37, no. 3, pp. 631-645, 2002.

[2] V. Liu, P. Kipnis, M. K. Gould, and G. J. Escobar, Length of stay predictions: Improvements through the use of automated laboratory and comorbidity variables, Medical care, vol. 48, no. 8, pp. 739 744, 2010.

[3] Panchami V U and N. Radhika,A Novel Approach for Predicting The Length of Hospital Stay with DBSCAN and supervised classification Algorithms, IEEE publisher,pp.207-212, 17-19 Feb,2014.

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

[5] E. K. Kulinskaya and H. D. Gao, Length of stay as a performance indicator: robust statistical methodology, IMA JOURNAL OF MANAGEMENT MATHEMATICS, vol. 16, no.4, pp. 369381, 2005.

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

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

[8] DBSCAN. Available at http://en.wikipedia.org/wiki/DBSCAN , accessed on 30th April, 2013.

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

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

[11] Bankers algorithm. Available at http://en.wikipedia.org/wiki/Banker'salgorithm", accessed on 14th November 2014.