Anomaly Detection System For Healthcare Resource Usage in Machine Learning

1Dr. J.M. Gnanasekar M.E.,Ph.D , 2Mr.C.Natarajan M.E (Ph.D)
1Professor , Dept. of Computer Science and Engineering,Sri Venkateswara College Of Engineering, Chennai
2Ph.D Scholar, Saveetha University, Chennai

ABSTRACT
Data mining approaches have been widely applied in the field of healthcare. Patient Medical Records contains vast clinical informations about patient conditions along with treatment and its procedure. Systematic healthcare utilization analysis use these observational datas to guide resource planning and improve the quality of care delivery while reducing cost. Here present a framework for utilization analysis that can be easily applied to the Diabetes population. The framework includes patients profiling with the disease entries and the patient conditions with treatment procedure, and contextual anomaly detection to provide the better healthcare delivery for the normal and abnormal patients by classifying the different kind of clinical characteristics patients in to clusters and form the patterns of disease evolution. Have to provide the corrective actions for the anomalies.

Keywords
Contextual anomaly,Clustering,Decision tree,Classification

INTRODUCTION
The Contextual anomaly detection is the process of finding the deviated patterns from the regular patterns. Every patients have different clinical characteristics, some of them are having similar characters, clustering is the one type of datamining functionalities, by using that we will form the cluster based up on similar characters and the anomalies will be filtered easily and placed in another cluster. The cluster will be formed such that the similarities between two cluster will be less and at the same time the similarities between the same cluster will be maximum. The framework includes two main components. The first component is patients profiling, with the disease entries and the patient conditions with treatment procedure, and the second component is Contextual Anomaly Detection for the efficient healthcare delivery. Typical anomaly detection methods identifying the data instances that deviate from the majority of the samples. A given patients profiling instance may be perfectly normal for one patient, but unexpected for another patient with different clinical conditions. We propose novel methods for contextual anomaly detection designed to detect anomalies. Our method is based on building models trained from observational data to compute the expected levels for each patient given his/her expected and actual levels based on well established statistical methods and provide the efficient medical treatment.

The patient healthcare utilization patterns analysis objectives include guiding resource planning, allocation and the medical needs and efficiency of health care service and procedures. This analysis improving the importance for health care institutions to ensure effective and efficient patient care delivery. Patient medical records include a large number of entries related to patient conditions along with treatments and procedures. Analysis based on such observational data collected and carried out in a systematic manner to improve care delivery in many ways. The anomaly detection is to identify the patterns that are usual and unusual clinical characteristics, including both normal and abnormal cases. This incurs unnecessary cost and waste of precious healthcare resources and ensures the way to provide efficient care delivery. The problems caused by the treatment without having such kind of analysis are very large and these cases to be avoided to increase the confidentiality of the medical treatment to the public. After finding the abnormal cases the corrective action will be taken.

RELATED WORK
Existing work on medical utilization pattern analysis has focused on disease specific studies and has not directly proposed a general framework for addressing the issues of anomaly detection. For example, a clustering method introduced by Barsky et al. to detect medical care utilization patterns for somatizing patients [2]. Nicholson et al. conducted research on patterns of ambulatory care use for gynecologic conditions. Eisele et al. studied the ambulatory medical care utilization patterns before and after the diagnosis of dementia in Germany [7]. Ruchlin et al. learned the resident medical care utilization patterns in continuing care retirement communities. Bushche et al. analyzed ambulatory medical care utilization by elderly patients in relation to patient conditions in Germany [5]. While these past studies each shed valuable light on the factors affecting the pattern of utilization in a specific disease condition, they were not designed to provide systematic approaches that can be adopted for finding the anomalies on any given patient population. Anomaly detection as a general topic has been studied for wide ranging domains including financial fraud detection, industrial damage detection, social media analysis, and medical and public health anomaly detection [6].

Utilization analyses are two main types of anomaly detection, namely Point Anomalies and Contextual Anomalies. Point anomalies refer to cases where an individual data instance (e.g., number of visits of different types) can be considered as anomalous to the rest of the data. The utilization profiling component of our proposed framework can be considered an instance of this type of anomaly detection using a classification based approach. In the classification based approach, singleton class as well as small clusters is considered. Based on total cost or a particular type of utilization (e.g. emergency visits), such multidimensional vector model analysis provides a more comprehensive understanding of the characteristics of different types of utilization.
In the second component apply the concept of contextual anomaly detection to this domain and develop an expectation modeling based approach to identify patients with anomalous utilization records. The basic idea is as follows. A regression model is trained from the observational data of recorded patients characteristics and corresponding utilization profiles. This model is then used to calculate the expected utilization level (behavior) with respect to any given patient profile (context). Then a comparison is made between the observed and expected behaviors and the Grubb’s test which is widely used for anomaly detection [10, 6] is deployed to determine whether there is an anomaly.

**UTILIZATION PROFILING**

Use \( p_i \) to indicate the \( i \)-th patient. The whole patient population set is denoted by \( P = \{ p_1, p_2, \cdots, p_n \} \), where \( n \) is the number of patients. The patient’s utilization is characterized by the number of different types of visits (e.g., visit to Primary Care Physician (PCP), visit to specialist, lab visit, etc.) incurred by this patient during a certain time period. At each step, two nearest clusters are merged. Here the distance between two clusters \( P_i \) and \( P_j \) is measured by

\[
d(P_i, P_j) = \frac{1}{n_i n_j} \sum_{p_i \in P_i} \sum_{p_j \in P_j} ||x_i - x_j||
\]

where \( n_i, n_j \) are the sizes of \( P_i, P_j \), and \( ||x_i - x_j|| = \sqrt{(x_i - x_j)\cdot(x_i - x_j)} \) is the Euclidean distance between \( x_i \) and \( x_j \). It can be easily seen that this is in fact the average distance between all pairs of data points with one in \( P_i \) and one in \( P_j \).

**Contextual Anomaly Detection**

Contextual anomaly detection approach consists of the following three steps. First, we learn functions that map clinical characteristics (contextual attributes) to utilization characteristics. These regression models are then used to estimate the expected number of visits of each type. Finally, a statistical test (Grubb’s test [10]) is applied to check if a significant difference exists between expected and actual utilization levels.

In the first step, the contextual attributes include patient demographics (age and gender) and clinical features characterized by ICD-9 codes. One potential shortcoming of using ICD-9 codes alone is that they do not adequately reflect clinical relations and risk groupings among different diagnoses. Much past work on Health Risk Assessment (HRA) has deployed various methods of generating risk groups, and reported improved accuracy of healthcare cost prediction using such groupings. We adopted the Hierarchical Condition Categories (HCC) used in Medicare Risk Adjustment provided by CMS (Centers for Medicare and Medicaid Services) in addition to the ICD-9 codes in the contextual attributes. For both ICD-9 codes and HCC codes, the clinical feature is defined as the percent of times that specific diagnosis was given in the utilization analysis period, which provides a measure of dominance of the corresponding condition for a patient. The target variables for the expectation models are the behavioral attributes, which in these cases are the numbers of visits for the different utilization types. A separate expectation model is built independently for each utilization type.

For the regression model, we explored several advanced function learners:

- **Classification And Regression Trees (CART)** [3] is similar to a decision tree except at the leaf level a regression model is constructed in order to map to a continuous target variable, instead of doing a majority vote as in a decision tree classifier.
- **Random Forest (RF)** [4] is an ensemble version of CART where multiple CART trees are built on the bootstrapping samples of the entire patient set. Here the bootstrapping is uniform such that each data point has an equal opportunity to be sampled with replacement.

**Multivariate Adaptive Regression Splines (MARS)** [8] (MARS) is a non-parametric regression technique and can be seen as an extension of linear models that automatically models non-linearities and interactions between variables. We used 10-fold cross-validation to evaluate these different methods. In our experiments, Random Forest consistently outperform all other methods in all cases, and was thus adopted in our system. Once the regression models have been trained, they are used to compute the expected level of utilization of each specific type for each patient given the contextual attributes of the patient. The difference between the expected and actual utilization (the residual error) can then be used to determine whether there is an anomaly. Intuitively, an actual utilization level should be declared anomalous if it deviates too much from the expected level. The key question to answer is: how much is too much? Certain utilization types may naturally have a wider range of variability associated with them than others and thus should be allowed larger deviation. We deploy Grubb’s test which had been widely used in the anomaly detection literature [10, 6] to take into consideration this inherent variability.

**Reduction Techniques – Principal Component Analysis**

The Reduction techniques are the data preprocessing techniques which involves giving the best suitable data’s to the mining purposes. The reduction is the process of changing the original data’s into reduced data without any loss. The original form of the data’s are changed to the new format with the reduced size. If the data’s are reduced to other forms and also the resultant data’s are equal to the original data when we are joining together then the reduction is called lossless reduction. If the data’s are lost while rejoining the tables, then that type of reduction is called lossy reduction.
Why we choose reduction before mining is, when the data grows very rapidly and if the big data involved means, without the reduction techniques we can’t give best pattern mining. Because we can’t easily observe and learn and predict any pattern from these large datasets. For that we used reduction techniques to reduce the size of the datasets and also that reduction to be a lossless reduction. There are five types of reduction techniques involved such as Data cube aggregation, Attribute subset selection, Dimensionality reduction, numerosity reduction and the last one is Data discretization and Concept hierarchy generation. Among these we chose Dimensionality reduction methods. Dimensionality reduction consists of two types. One is Discrete Wavelet Transforms, and another one is Principal Components Analysis. Among these two for our medical recorded data, the later one is best suited. And also it will provide good result too. Principal Component Analysis is the reduction method having four basic procedures as follows.

* First have to change the input records to normalized data.
* Form the orthogonal vector for the above normalized data. The name of the vector is called as Principal Components.
* Have to arrange the principal components in Decreasing of most significant order of the vectors.
* After arranging the decreasing of order of most significant first, the most insignificant data vectors will be deleted. Thus the necessary data vectors will select and thus the large data sets get reduced into adaptable size.

Here the following figure shows how clustering formed.

![Hierarchical Clustering](image)

**Figure 1:** Hierarchical Clustering

The patient population can be segmented in to 2 steps. First one is data segmentation and form the micro clusters. And the second step is forming the hierarchical clustering.

**RESULTS AND ANALYSIS: DIABETES PATIENT MANAGEMENT**

**Data Description**

The proposed framework has been tested using claims data collected from a network of physicians over a one year period. While the framework is very general and can be applied to any patient population, it is useful to focus on a specific use case to investigate whether the results are clinically meaningful. We thus constrained our experiments to the diabetes patient population which contains a total of 7,667 patients. For this population 98% of the total visits belong to one of the top six visit types as given in Table 1, which also provides statistics for these top visit categories.

**Table 1:** Statistics and Descriptions of the Patients Visits

<table>
<thead>
<tr>
<th>Type of Visits</th>
<th>Description of the Visits</th>
<th>Total No.of Visits of the patients</th>
<th>Percentiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PCP visit</td>
<td>61,253</td>
<td>6 12 20</td>
</tr>
<tr>
<td>2</td>
<td>Specialist visit</td>
<td>77,255</td>
<td>6 15 32</td>
</tr>
<tr>
<td>3</td>
<td>Emergency visit</td>
<td>5,731</td>
<td>0 0 4</td>
</tr>
<tr>
<td>4</td>
<td>Outpatient hospital visits</td>
<td>34,047</td>
<td>0 6 18</td>
</tr>
<tr>
<td>5</td>
<td>Inpatient hospital visits</td>
<td>20,826</td>
<td>0 0 14</td>
</tr>
<tr>
<td>6</td>
<td>Patient’s home</td>
<td>15,389</td>
<td>0 4 9</td>
</tr>
</tbody>
</table>
From the table it can be clearly observed that the majority of the patients had relatively low level of utilization. For example, the 50 percentile of the total number of visits is only 12, i.e., half of the patients only made up to 12 visits to medical facilities during the year.

Results for Utilization Segmentation and Hot Spotting

The modified Hierarchical Agglomerative Clustering approach described was applied to this patient population. The resulting dendrogram can then be explored interactively by a domain expert. For each cluster, the user can examine the representative utilization profile of the cluster (computed as the cluster mean), average cost, and patient characteristics such as mean age, sex ratio and dominant diagnoses. For this diabetes population, a close examination by the MD in our group revealed that a total of 10 clusters provide a meaningful level of segmentation. Hot spotting are the patients who visit the doctor frequently or rarely. Hot spotting can then be performed by analyzing small and isolated high utilization groups. The second component is Contextual Anomaly Detection for Utilization. Typical anomaly detection methods focus on identifying data instances that deviate from the majority of the samples.

The following table 2 shows the representative utilization profile for each cluster (cluster mean). It shows for each cluster the cluster size, average cost, average age and a clinical description of the cluster derived by the MD based on information provided by the system as explained above. It can be clearly seen from this analysis that clusters 1-4 represent well managed patients with varying but stable conditions, leading to relatively low level of utilization and cost. Clusters 5, 6 and 8 represent patients with more advanced disease state and advancing complications, thus requiring increased utilization. Finally, clusters 7, 9 and 10 are the “hot spot” patients with advanced conditions requiring intense utilization of different types.

These are patients who will likely benefit from an intensive disease management program. The maximum number of patients from the cluster High-utilizing diabetics with advancing renal failure, cardiac and vascular disease. These patients require frequent use of specialists. This shows that the results of the utilization segmentation. The hot spotting patients are the patients with Older age that is above 70 having diabetics likely to have a high percentage of smokers with costly complications, COPD and complex outpatient treatments. The second categories are the age of above 55 and having End-stage diabetics with advanced complications including renal failure, heart failure, poorly-controlled hypertension, requiring frequent hospitalizations, outpatient hospital procedures, and home health visits.

<table>
<thead>
<tr>
<th>Cluster Id</th>
<th>Size</th>
<th>Average cost per patient</th>
<th>Average Age</th>
<th>Description of Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>377</td>
<td>5758</td>
<td>69</td>
<td>Diabetics with Hypertension, cardiac arrhythmias, with Specialty Care and Outpatient Hospital Clinics, ER Visits</td>
</tr>
<tr>
<td>2</td>
<td>959</td>
<td>4720</td>
<td>70</td>
<td>Diabetics with Hypertension, Hyperlipidemia and some cardiac disease with requiring some Specialist visits</td>
</tr>
<tr>
<td>3</td>
<td>807</td>
<td>2580</td>
<td>66</td>
<td>Diabetics with complications of Hypertension, Hyperlipidemia and some with cardiac arrhythmias, avoiding Specialist, Outpatient Clinics.</td>
</tr>
<tr>
<td>4</td>
<td>5013</td>
<td>1573</td>
<td>63</td>
<td>younger patients with uncomplicated Diabetics with Hypertension, Hyperlipidemia, making minimal use of services.</td>
</tr>
<tr>
<td>5</td>
<td>239</td>
<td>10150</td>
<td>69</td>
<td>Diabetics with Hypertension, cardiac arrhythmias, and arthritis, making extensive use of Specialists, while avoiding Hospitalizations</td>
</tr>
<tr>
<td>6</td>
<td>127</td>
<td>11,738</td>
<td>69</td>
<td>Diabetics with increasing comorbidities, and complications requiring periodic hospitalization for exacerbations of heart failure.</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>66,480</td>
<td>63</td>
<td>High-utilizing diabetics with advancing renal failure, cardiac and vascular disease. These patients require frequent use of specialists</td>
</tr>
<tr>
<td>8</td>
<td>112</td>
<td>16,375</td>
<td>66</td>
<td>Diabetics with advancing complications being managed by specialists. Higher representation of women.</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>42,559</td>
<td>71</td>
<td>Older diabetics likely to have a high percentage of smokers with costly complications, COPD and complex outpatient treatment.</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
<td>41,980</td>
<td>57</td>
<td>End-stage diabetics with advanced complications including heart failure, un-controlled hypertension, requiring frequent hospitalizations.</td>
</tr>
</tbody>
</table>
Results for Contextual Anomaly Detection

A separate expectation model was trained for each one of the top six utilization types using Random Forest regression model, using diagnoses, age and sex as the contextual attributes to predict the expected level of utilization. Table shows the prediction results for each one of the six utilization types using standard 10 fold cross validation. As can be seen in the table, a positive $R^2$ measure was achieved for all utilization types, including even Emergency visit, which is particularly difficult to predict because of the sparsity of the event, and large degree of randomness (e.g., accidents). For visit type involving less degree of randomness, the performance improves as expected. Particularly, for Specialist and Inpatient hospital visits we achieved $R^2$ values greater than 0.3. These results indicate that the proposed expectation model can indeed lead to better prediction of expected utilization level than using population mean, which should lead to more personalized and clinically meaningful anomaly detection. For each patient the difference between the expected level and actual level of utilization is compared against the mean residual error and the Grubb’s test is used to determine if this different is anomalous. A patient is considered anomalous if he/she is signaled as such for at least one of the utilization types. Using a significance level of 0.05 in Grubb’s test, a total of 51 anomalies were detected. These anomalies can then be explored in the system by examining the actual vs. expected utilization levels, and contextual attributes including age, sex and dominant diagnosis to determine the next step of investigation. Here we provide sample investigations of three of the patients with anomalous utilizations. Figure shows the expected vs. actual utilization for each patient, and Table provides the characteristics.

CONCLUSIONS

The framework for utilization analysis that can be used to perform systematic and timely identifications of heavy users of different types as well as contextual anomalies, i.e., utilization instances that are unexpected given patients’ clinical characteristics. In order to assess the general applicability of the framework, in this initial exploration we restricted our experiments and analysis to the most widely available type of data, i.e., claims data including diagnosis, demographics, and medical utilization records. Our evaluations and case studies demonstrate the usefulness of the proposed approaches in identifying clinically meaningful instances for both hot spotting and anomaly detection, using the most basic observational data as described above. Clearly many other data sources such as EMRs and patient and disease registries could provide additional information relevant to utilization analysis. In our future work we plan to expand our framework to leverage these additional data sources to provide enhanced performance and additional actionable insight. Another limitation of the proposed methods is that we currently do not consider temporal relationships among different medical events or encounters.

REFERENCES

Author’s biography

Dr. J. M. Gnanasekar is working as a Professor in Department of Computer Science and Engineering, Sri Venkateswara College of Engineering (SVCE), Sriperumpudur, Kancheepuram (District), TN, India. He completed his ME in Computer Science and Engineering and PhD from College of Engineering, Guindy, Anna University, Chennai. His areas of interest include Wireless sensor networks, security, data mining, NLP and cloud computing. He has more than 20 years of experience in teaching and research in the area of computer science and engineering.

Natarajan Chellasamy M.E(Ph.D) was born in Virudhunagar District of Tamil nadu. He had finished his B.E degree in Computer Science and Engineering in Manonmanium Sundaranar University and he had finished his M.E degree in Computer Science and Engineering in Anna University, Chennai. He is doing his Ph.D in Computer Science and Engineering in Saveetha University since June 2013. His area of Interests are data mining, Neural networks.

This work is licensed under a Creative Commons Attribution 4.0 International License.

DOI: 10.24297/ijclds.v5i1.5908