Utilization of Data- Mining Techniques for Evaluation of Patterns of Asthma Drugs Use by Ambulatory Patients in a Large Health Maintenance Organization

Mark Last
Ben-Gurion University of the Negev, Israel
mlast@bgu.ac.il

Rafael Carel
University of Haifa, Israel
rcarel@research.haifa.ac.il

Dotan Barak
Ben-Gurion University of the Negev, Israel
dotanbarak@hotmail.com

Abstract

A major problem of drugs utilization is to identify outlier patients who are using large quantities of drugs over extended periods of time. Today, healthcare and health insurance systems have to deal with an increased number of patients suffering from chronic diseases, such as asthma, who are continuously using a combination of several medications. This has caused a substantial increase in the cost of providing healthcare for such patients. In Israel, 11% of the national health care budget is spent on medications. However, healthcare management operations do not have the information that can assist in determining whether extensive multi-year drug utilization by a chronic patient is an outlier or misuse of resources. In this work, we construct a prediction model for asthma drug utilization by applying novel methods of knowledge discovery in time-series databases to a multi-year asthma drug utilization data set. Methods of mining utilization patterns combine clustering algorithms, clustering validity measures, and decision-tree classification algorithms. This methodology is applied to a regional patients' database maintained in 'Clalit Health Services' HMO, Beer-Sheva, Israel between January 2000 and November 2002. The clustering results reveal that 274 asthma patients who received 9,319 prescriptions during that period can be partitioned into three groups of utilization patterns, where ten patients (3.6%) who used 1,333 prescriptions (14.3%) are classified as outliers. The classification results show that the use of corticosteroids medications (oral or by inhalation) and the age of a patient can be considered as the main predictive factors in the induced models.

1. Introduction

Asthma is a chronic inflammatory disease of airways that is characterized by increased responsiveness of the tracheo-bronchial tree to a multiplicity of stimuli. It is an episodic illness with acute exacerbations interspaced with symptom-free periods [10]. Most of the care for adult asthma patients is provided by primary health care physicians. It is estimated that about 6% of the adult population in the USA are diagnosed as suffering from asthma [15]. Optimal management of the disease requires continuous medical surveillance with periodical adaptation of the treatment regimen to the clinical state of the patients. The disease has significant clinical and economic impacts both on patients and the health care system. As a rule, many of the patients are using simultaneously a combination of drugs for extended periods of time in order to control their disease. Thus, the costs involved in care of these patients are substantial.

Currently, in many ambulatory health care systems, the extensive use of computers and computerized data-bases (DB) for routine services is quite common. In addition, medical information systems (MIS) and medical decision support system (MDSS) are also used in order to standardize and improve the quality of health care.

Data mining (DM) is the core stage of knowledge discovery in databases (KDD), which is a "non-trivial extraction of implicit, novel, and potentially useful information from data" [6]. It applies machine learning and statistical methods in order to discover areas of previously unknown knowledge. As a rule, the KDD process involves the following steps: data selection, data pre-processing, transformation, DM (induction of useful patterns), and interpretation of results.

Two common data mining tasks are: (a) *cluster analysis* aimed at organizing a given dataset into groups (clusters) of similar objects or characteristics and (b) *classification* aimed at predicting the class of objects whose class label is unknown. One of the most common classification models is the *decision tree*, which is a tree-like structure where each internal

0-7695-3019-2/07 $25.00 © 2007 IEEE
DOI 10.1109/ICDMW.2007.50

node denotes a test on a predictive attribute and each branch denotes an attribute value. A leaf node represents predicted classes or class distributions [11]. An unlabeled object is classified by starting at the topmost (root) node of the tree, then traversing the tree, based on the values of the predictive attributes in this object. Decision-tree techniques assume that the data objects are described by a fixed set of attributes, where each predictive attribute takes a small number of disjoint possible values and the target (dependent) variable has discrete output values, each value representing a class label.

There are several known algorithms of decision-tree induction: ID3 - which uses information gain with statistical pre-pruning, C4.5, an advanced version of ID3, and probably the most popular decision-tree algorithm [13], CART, which minimizes a cost-complexity function, See5 - which builds several models and uses unequal misclassification costs, and IFN – Info-Fuzzy Network which utilizes information theory to minimize the number of predictive attributes in a decision-tree model [7] [9]. In [7], the IFN algorithm is shown empirically to produce more compact models than C4.5, while preserving nearly the same level of classification accuracy.

Cluster analysis or clustering, is a data mining technique aimed at grouping the data objects into classes or clusters so that objects within one cluster have high similarity to each other, but are very dissimilar to objects in all other clusters [6]. Clustering is a form of *unsupervised learning*, since it does not rely on class-labeled training examples. One of the most popular partitioning clustering methods is *k-means*, which splits a set of *n* objects into *k* clusters. Several methods for determining the optimal number of clusters *k* have been proposed [5].

A time series database (TSDB) may include various types of variables with at least one temporal dimension. A large portion of today's databases, especially in financial, scientific, and medical domains, can be considered as time series data. Recently there has been a growing interest in mining time series databases, with attempts to cluster, classify [8], maintain, and index temporal data .

Medical applications of data mining include prediction of the effectiveness of surgical procedures, medical tests, and medications as well as discovery of relationships between clinical and pathological data [12]. Clinical databases store large amounts of information about patients and their medical conditions. Data mining techniques applied to these databases discover relationships and patterns which are helpful in studying progression and management of diseases.

In this paper, we apply data mining techniques for discovery and evaluation of patterns of asthma medication usage by ambulatory patients of a large health maintenance organization. The rest of our paper is organized as follows: Section 2 describes the process of data acquisition for this study and indicates the pre-processing operations applied to the obtained data. In Section 3, we explore the resulting 35-months asthma drug utilization dataset using clustering and classification techniques. In Section 4, we conclude our study with the summary of main results.

2. Data Preparation

2.1 Dataset Acquisition

Several computerized databases are used by the Clalit HMO (Health Maintenance Organization) in Southern Israel. In this study, we used the Pharmacy Registry (PR) as one of the sources for the data mining process. All medications prescribed and given to patients in any pharmacy in the region are recorded in one central (regional) PR, primarily for administrative and financial control. For each prescription the following data is recorded: date, patient name, his/her unique ID number (similar to the Social Security Number in the US), name and code of the prescribing physician, date when the prescription was issued, name and quantity of the prescribed drug, and code of the clinic where the patient is registered. We have extracted from the PR a data set of all asthma-related prescriptions issued to adult patients (ages 25-65) registered in 6 clinics during the period of January 2000 to November 2002.

Information regarding the diagnoses of these individuals was extracted from their computerized medical records in the 6 clinics of interest. Clalit Health Services is managing its patients' electronic medical records (EMR) in a system known as Clicks (Computerized Personal File). The Clicks files contain demographic data, records of each visit to the clinic, and a list of diagnoses relevant to the patient in question. In our study, ICD-9 codes of each record were used to identify asthma patients.

Joining the two databases (the Pharmacy Registry and the Clicks) on the unique patient ID number (attribute) allowed us to create a data set of asthma patients along with the type and amount of asthma medications that they have received during the study period. The drug prescriptions TSDB (Time-Series-DB) that was extracted from the PR database initially included 11,312 records of asthma drugs prescribed for patients of the six clinics included in our study.

In these clinics, we found about 2,100 records of patients with diagnosis of asthma (in the age group of 25 to 65).

2.2 Data Pre-Processing

In the data cleaning stage, duplicate records were removed and missing attributes completed by using appropriate data (e.g. physician's number and name could be restored using several records of a drug data set for the same patient). Records with missing drug name and/or date attributes were completely removed from the dataset.

We had to perform a profound data cleaning on the ID attribute because of the non-uniform way patient IDs are recorded in the two databases. The Israel Personal ID numbers include eight digits plus an audit digit. We used the Israel Ministry of Interior algorithm for converting the incomplete ID numbers from the eight-digit to the nine-digit format.

Based on recommendations made by an asthma specialist, we decided to consider a patient as an "asthmatic" if she/he received five or more prescriptions of medications commonly used for treating asthma during the three-year period of study. The number of asthma patients remained in the drug data set after applying this restriction was 3,144. While checking for duplicate records and missing attributes, additional records were discarded leaving us with 2,801 patients in the drug data set. In the dataset of asthma patients, we were left with only 2,104 records after similar pre-processing operations.

Since some asthma drugs were seldom used, we aggregated all drugs according to their pharmacological structure. The four types with the highest utilization were: Beta2 agonists, Corticosteroids, Anticholinergic, and Xanthines. Based on the patient IDs and the distribution dates, we aggregated all prescriptions to monthly utilization data for each of the 35 months in the period of study.

After joining the medication use and the patient tables on the correct IDs, we arrived at only 308 asthma patients using the four most common drug types. Subsequently, we have found that some patients appeared more than once in the dataset because they were included in the lists of several clinics. Eventually 274 patients with asthma diagnosis and more than five records of prescriptions were identified for our next stage of mining 35 months of drug utilization. Using a lookup table, all 274 real ID's were replaced with a serial number from 1 to 274, due to privacy issues.

3. Mining Drug Utilization Data

3.1 Clustering Patients by Drug Utilization Data

We have used the *K*-means clustering algorithm to identify groups of patients having similar utilization patterns of asthma drugs over the period of 35 months. The features we used to characterize each patient were the number of prescriptions issued per each month. All records (time series) had 35 monthly values for drug utilization; hence, we clustered vectors having 35 dimensions. The overall average utilization was 0.97 prescriptions per month, but as shown in Figure 1, the average monthly drug utilization sorted in ascending order is similar to a log-like function. About 90% (249) of patients consume less than two prescriptions per month.

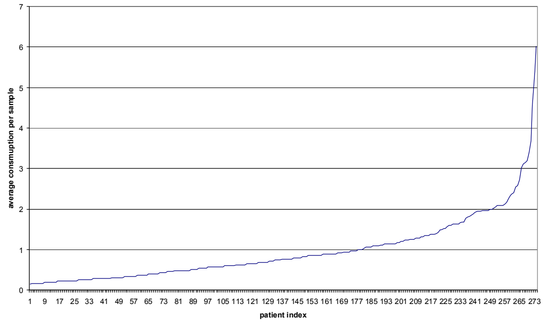

Figure 1. Average Monthly Utilization per Patient (sorted in ascending order)

For each object, the *K*-means algorithm locates iteratively the closest centroid that optimizes the quality of the clustering. In each iteration, objects are reallocated to the nearest cluster centroid. When no reallocation of the objects in any of the clusters takes place, the process terminates. Since physicians were unable to indicate the correct number of patient groups, we used *k*=2 to k= 10 values with the purpose of finding the optimal number of clusters. We calculated the inter-cluster and the intra-cluster dissimilarity using the Euclidean distance between two vectors.

Clustering validation indices measure how well a dataset is clustered with different settings. In our study, we used the following validation techniques to determine the best number of clusters *k*: Dunn Validity Index and Silhouette Validation Method. Both techniques are briefly described below.

Dunn Validity Index ([2] [4]). This index is seeking for compact and well separated clusters. For any partition of clusters, where c_i and c_j are i and j

clusters of such partition, the Dunn parameter *D* is calculated by the following equation:

$$D = \min_{1 \le i \le n} \left\{ \min_{\substack{1 \le j \le n \\ i \ne j}} \left\{ \frac{d(c_i, c_j)}{\max_{1 \le k \le n} \left\{ d'(c_k) \right\}} \right\} \right\} \qquad (1)$$

In Eq. (1) above, *d* (c_i, c_j) is the distance between clusters *i* and *j* (inter-cluster); $d'(c_k)$ is the distance between objects in cluster *k* (intra-cluster); and *n* is the total number of clusters. Eq. (1) is aimed at minimizing the intra-cluster distance and maximizing the inter-cluster distance. Thus, the best number of clusters should maximize this equation.

Silhouette Validation Method ([2] [14]). In this method, each cluster is represented by a type of silhouette, which is the ratio between its compactness and distribution. The technique computes the silhouette width for every object, the average width for each cluster, and the average silhouette width for the entire dataset.

The formula we use to construct the object silhouette is:

$$S(i) = \frac{(b(i) - a(i))}{\max\{a(i), b(i)\}} \qquad (2)$$

Where *a(i)* is the average distance (dissimilarity) of an object *i* to all other objects in its cluster and *b(i)* is the average distance of an object *i* to all other objects in the closest cluster.

Based on Eq. (2), $-1<S(i)<1$. If $b(i)>>a(i)$ then $S(i) \rightarrow 1$, the object is clustered as best as possible. If $b(i)=a(i)$ then $S(i) = 0$, the object is indifferent and can be allocated in another cluster. if $b(i)<<a(i)$ then $S(i) \rightarrow -1$, the object is misclassified and located between the clusters, with a high tendency to the other cluster. The average of all silhouette widths for all objects in the dataset is the average *S(i)* for all items. The optimal number of silhouette widths is taken as the largest *S(i)*, which indicates the best clustering. The main idea of both indices (Dunn and Silhouette) is that the inter-similarity will be as small as possible and the intra-similarity large but while Dunn's method uses the inter-similarity per cluster, the Silhouette method constructs its formula by inter-cluster similarity per object.

We performed ten clustering runs for each value of *k*; each run started with different objects as initial centroids. In each run, we calculated the objects in the clusters, the centroid of each cluster, the average inter-cluster similarity, the intra-cluster similarity of each object inside its cluster, the average intra-cluster similarity of each cluster, the Dunn index for each run, the Silhouette index of each object, and, finally, the total Silhouette index of each run.

The results of two clusters and ten clusters were considered poor by the domain experts. For two clusters, all patients were trivially partitioned into two groups of high and low utilization, which has not provided any new insight into the distribution of drug usage. For ten clusters, in almost every run we obtained a cluster of one object, which made us believe that the actual number of multi-object clusters is not greater than nine. As can be seen in Figure 2 both validation indices tend to become smaller as *k* is increased. Thus, we have concluded that the best number of clusters for this dataset should be three.

The results from this clustering have shown that the utilization time series (representing asthma patients) should be partitioned into the following three clusters: ten patients (3.6%) with very high utilization (denoted as Cluster 0), 83 patients (30.3%) with medium utilization (denoted as Cluster 1), and 181 patients (66.1%) with low utilization (denoted as Cluster 2). The average monthly number of prescriptions was 3.81, 1.57, and 0.53 for patients in Clusters 0, 1, and 2, respectively. Although the clusters were constructed from vectors of 35 months, the cluster assignment of all objects but one matched their average monthly utilization. The most interesting cluster is the smallest one for obvious reasons. In addition, the largest cluster shows that on average an asthmatic patient gets about one prescription per month, while about 50% of patients in this cluster get only one prescription every two months or even less than that. In the next sub-section, we are applying classification algorithms to these clusters with the purpose of discovering the most influential features that affect the monthly drug utilization for each patient.

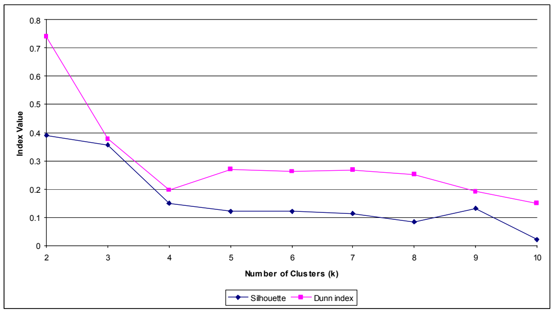

Figure 2. Dunn and Silhouette Indices as a Function of *k*

3.2 Classification of Clustered Patients

In the second stage, after we have found the best clustering and assigned each patient in our dataset to her/his cluster, we can apply supervised classification algorithms to our data and thus identify the features and the models that can predict the drug utilization of each patient. The available features are sex, age, month of birth, physician code, clinic code, and average utilization of four drug types - Xanthine, Anticholinergic, Corticosteroids and Beta2_agonists. In our dataset, we have 274 patients in the ages between 26 and 67 from both genders, 39 physicians, and six clinics. The value in the column of every drug type is the percentage of usage of the specific drug type out of all drug types prescribed to the patient.

The summary of IFN results is shown in Table 1. The first row corresponds to the default confidence level of the algorithm (99.9%). This level is supposed to produce the most compact model resulting in the smallest number of predictive attributes. Two lower confidence levels (99.0% and 95.0%) can reveal additional or alternative explanatory features. The testing error rates were estimated using a hold-out set of 94 records. Based on the testing error, the first model, obtained with the highest confidence level of 99.9%, is the most accurate one though the difference vs. CL=0.95 appears to be negligible. In fact, the error rate of the 99.9% model (36.2%) is equal to the error rate of the majority rule applied to the testing records, which means that the induced model is not particularly useful for predicting the amount of drugs prescribed to a given patient though it does reveal the major factors related to drug utilization.

Table 1. Summary of IFN results for Three Confidence Levels

Conf. Level	Number of Nodes	Attributes in Descending Importance	Testing Error Rate
0.999	8	Corticosteroids, Age, Month of Birth	0.362
0.99	9	Corticosteroids, Beta2_agonists, Month of Birth	0.415
0.95	85	Corticosteroids, Age, Anticholinergic, Beta2_agonists, Sex	0.372

Some information-theoretic rules for the 99.9% confidence level appear in Table 2. As explained in [9], the rules having the highest positive weights are the most informative ones. Predicted clusters 0, 1, and 2 correspond to high, average, and low utilization, respectively. As we can see, the Corticosteroids Percentage is the best explanatory attribute. For example, Rule 4 shows that if Corticosteroids is more than 57% then the predicted Cluster is 2 (low utilization). When Corticosteroids is between 5% and 57%, the age of a patient and sometimes even his/her Month of Birth affects the predicted cluster as well.

Table 2. IFN Rules for CL = 0.999

Rule No.	Rule text	Weight
1	If Corticosteroids is between 0 and 0.05 then Cluster is not 1	-0.0306
2	If Corticosteroids is between 0 and 0.05 then Cluster is 2	0.0661
3	If Corticosteroids is more than 0.57 then Cluster is not 1	-0.0197
4	If Corticosteroids is more than 0.57 then Cluster is 2	0.0792
5	If Corticosteroids is between 0.05 and 0.57 and Age is between 26 and 55 then Cluster is not 0	-0.0018
6	If Corticosteroids is between 0.05 and 0.57 and Age is between 26 and 55 then Cluster is 1	0.0007

We applied the C4.5 algorithm to the same dataset with the default CL value of 0.25. The maximal tree of the size of 131 nodes was pruned to a smaller one with 7 nodes (see Table 3). Similar to IFN, the C4.5 decision tree (after pruning) considers Corticosteroids as the most influential attribute. Additional predictive attributes include Beta2_agonists utilization, while the attribute with least influence is the patient age. According to the simplified C4.5 decision tree, no record is assigned to Cluster 0, which is the smallest cluster of high-usage patients. The testing error rate of the simplified tree was estimated using the same hold-out set of 94 records as in IFN runs. The testing error rate of the simplified C4.5 decision tree is equal to the IFN error rate with CL=0.999: 36.2%.

Table 3 C4.5 Decision Tree (after pruning)

```
Corticosteroids > 0.56 : 2 (43.0/3.8)
Corticosteroids <= 0.56 :
|  Beta2_agonists > 0.94 : 2 (42.0/4.9)
|  Beta2_agonists <= 0.94 :
|  |  Age <= 61 : 2 (144.0/63.6)
|  |  Age > 61 : 1 (45.0/20.8)
```

A comparison of IFN and C4.5 shows that both algorithms produce classification models of similar accuracy for this data. Additionally, both algorithms treat the corticosteroids drug type as the most influential attribute. Beta2_agonists are considered as the second significant attribute by most models, while Age, Month of Birth, Anticholinergic, and Gender appear to have a minor influence only. The apparent correlation between the patient age and the amount of drugs used by asthma patients is not well documented in the medical literature, although older patients may suffer from more than one chronic disease, which may increase their health awareness. A potential relationship between month of birth and asthma was suggested in previous epidemiological studies [2].

4. Conclusions

In our study, we have partitioned asthma patients into three groups of: high usage (about 3.6% of the patients), medium usage (30.3%), and low usage (66.1% of the patients). Two classification algorithms (IFN and C4.5) were applied to the clustered data. According to both classification algorithms, the corticosteroids (both oral and by inhalation) is the major predictor. Minor prediction factors that were identified are: Age, Anticholinergic drugs, Month of Birth, Beta2-agonists, Gender (according to IFN) and Beta2-agonists, Xanthine, Corticosteroids (according to C4.5).

Application of data mining methods can help physicians and health management organizations to monitor efficiently the utilization of drugs by chronic patients. Automated identification of patients at risk of suboptimal treatment or over utilization can help a physician in better controlling such patients.

A major limitation of this study is the sample size. Though we had about 88,000 asthma drug prescriptions, we could use only about 11% of all data. Another limitation is related to the quality of the data. Since the ID number was not uniform in all data sets, it caused some data loss.

Future research may focus on applying spatio-temporal data mining methods in order to find geographical patterns of asthma prevalence. The geographical area of interest may be in any size from a small neighborhood to the entire country.

Our study does not deal at all with utilization of specific drugs due to insufficient amount of data. Since many drugs have been dropped, further research with the same methods and using a complete HMO data set can explore a specific drug or drug types as treatment and prevention.

5. References

[1] R. Agrawal, C. Faloutsos, A. Swami, "Efficient Similarity Search in Sequence Databases", In *Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms*, Springer-Verlag, 1993: 69-84.

[2] H.R. Anderson, P.A. Bailey, J.M. Bland, "The Effect of Birth Month on Asthma, Eczema, Hayfever, Respiratory Symptoms, Lung Function, and Hospital Admissions for Asthma", Int J Epidemiol. 1981 Mar; 10(1): 45-51.

[3] F. Azuaje, "A Cluster Validity Framework for Genome Expression Data", *Bioinformatics* 2002; 18: 319-320.

[4] J.C. Dunn, "Well Separated Clusters and Optimal Fuzzy Partitions", *J. Cybern*. 1974; 4: 95-104.

[5] M. Halkidi, Y. Batistakis, M. Vazirgiannis, "On Clustering Validation Techniques", *J. Intell. Inf. Syst.* 2001; 17, 2-3: 107-145.

[6] J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, 2nd Edition, Morgan Kaufmann, 2006.

[7] M. Last and O. Maimon, "A Compact and Accurate Model for Classification", *IEEE Transactions on Knowledge and Data Engineering* 2004; 16, 2: 203-215.

[8] M. Last, Y. Klein, A. Kandel, "Knowledge Discovery in Time Series Databases", *IEEE Transactions on Systems, Man, and Cybernetics* 2001; 31, 1: 160-169.

[9] O. Maimon and M. Last, *Knowledge Discovery and Data Mining – The Info-Fuzzy Network (IFN) Methodology*, Kluwer Academic Publishers, Massive Computing, Boston, December 2000.

[10] E.R. McFadden, Asthma. Chapter 236, in *Harrison's Principles of Internal Medicine*. 16th ed. McGraw Hill, 2005.

[11] T. Mitchell, *Machine Learning*, McGraw Hill, 1997.

[12] J.C. Prather, D.F. Lobach, L.K. Goodwin, J.W. Hales, M.L. Hage, W.E. Hammond, "Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse", *Proc AMIA Annu Fall Symp*. 1997:101-5.

[13] J.R. Quinlan: *C4.5, Programs for Machine Learning*, Morgan Kaufmann, 1993.

[14] P.J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis" *Journal of Computational and Applied Mathematics* 1987; 20: 53-65.

[15] World Health Organization: http://www.who.int/mediacentre/factsheets/fs206/en/