JCBi JOURNAL OF CLINICAL BIOINFORMATICS

METHODOLOGY Open Access

Potential identification of pediatric asthma patients within pediatric research database using low rank matrix decomposition

Teeradache Viangteeravat

Abstract

Asthma is a prevalent disease in pediatric patients and most of the cases begin at very early years of life in children. Early identification of patients at high risk of developing the disease can alert us to provide them the best treatment to manage asthma symptoms. Often evaluating patients with high risk of developing asthma from huge data sets (e.g., electronic medical record) is challenging and very time consuming, and lack of complex analysis of data or proper clinical logic determination might produce invalid results and irrelevant treatments. In this article, we used data from the Pediatric Research Database (PRD) to develop an asthma prediction model from past All Patient Refined Diagnosis Related Groupings (APR-DRGs) coding assignments. The knowledge gleamed in this asthma prediction model, from both routinely use by physicians and experimental findings, will become fused into a knowledge-based database for dissemination to those involved with asthma patients. Success with this model may lead to expansion with other diseases.

Keywords: Clinical research, Translational research, Medical informatics, Biomedical informatics, Machine learning, Data mining, Feature extraction, Classification

Background

Data mining in medical informatics

Because of their predictive power, various healthcare systems are attempting to use available data mining techniques to discover hidden relationships as well as trends in huge data available within the clinical database and convert it to valuable information that can be used by physicians and other clinical decision markers. In general, data mining techniques can learn from what was happened in past examples and model oftentimes nonlinear relationships between independent and dependent variables. The resulting model provides formalized knowledge and prediction of outcome. For example, Shekar et al. used data mining based decision tree algorithm to discover the most common refractive error in both male and female [1]. Palaniappan et al. presented a prototype that combines the strengths of both an online analytical processing (OLAP) and data mining techniques for clinical decision support systems (DSS) [2]. Jonathan et al. used data mining techniques to explore the factors contributing to cost of prenatal care and outcomes [3]. Chae et al. used data mining approach analysis in health insurance domain [4]. With advanced data mining techniques to help evaluate healthcare utilization costs for employees and dependents in organizations [5].

More advanced machine learning methods, such as artificial neural networks and support vector machines, have been adopted to use in various areas of biomedical and bioinformatics, including genomics and proteomics [6]. For biological data, clustering is probably the most widely used data mining technique, such as clustering analysis, hierarchical clustering, *k*-means clustering, backpropagation neural networks, self-organization maps, fuzzy clustering, expectation maximization, and support vector machines [7,8]. Bayesian models were widely used to classify data into predefined classes based on a set of features. Given the training examples, a Bayesian model stores the probability of each class, the probability of each feature, and the probability of each feature given each class. When a new unseen example occurred, it can be

Correspondence: tviangte@uthsc.edu
Biomedical Informatics Core, Children's Foundation Research Institute, Department of Pediatrics, The University of Tennessee Health Science Center, 50 N. Dunlap, 38013, Memphis, TN, USA

classified according to these probabilities [9,10]. This classification technique is one of the most widely used in medical data mining. Decision tree models, such as the Iterative Dichotomiser 3 (ID3) Heuristic techniques belong to the subfield of machine learning. The ID3 Heuristic uses a technique called "entropy" to measure disorder in a set of data [11,12]. The idea behind the ID3 Heuristic is to find the best attribute to classify the records in the data set. The outcome is learned rules and a model used to predict unseen examples based on past seen examples. Non-negative matrix factorization (NMF) has been widely used in the field of text mining applications [13,14]. The only constraint that is unique from other methods is factorization of two matrices *W* and *H* from *V* (i.e., *nmf* (*V*) → *WH*) must be non-negative or all elements must be equal to or greater than zero. Typically, *W* and *H* are initialized with random non-negative values to start the NMF algorithm. The convergent time is varied and local minimum is not guaranteed [15].

Here, we are working on a methodology and classification technique in data mining called Low Rank Matrix Decomposition (LRMD) to allow computer to learn from what has happened in the past APR-DRGs datasets for asthma, able to extract dominant features, and then predict outcomes. The summary of APR-DRGs and the mathematics behind LRMD is discussed further below.

All patient refined diagnosis related groups (APR-DRGs)

APR-DRG is a grouping methodology developed in a joint effort between 3M Health Information Systems (HIS) and National Association of Children's Hospitals and Related Institutions (NACHRI). APR-DRGs are proprietary and have the most comprehensive and complete classification of any severity of illness system for pediatric patients. It was designed to be more appropriate for general population patients than the old Diagnosis Related Group (DRG) [16]. While the DRG was designed and normed on Medicare patients only, the APR-DRG was designed and normed on a general population. We use APR-DRG based weights normed on a pediatric patient population. There are 316 APR-DRGs, such common APR-DRG codes include but not limited to 138 Bronchiolitis/RSV pneumonia, 141 Asthma, 160 Major repair of heart anomaly, 225 Appendectomy, 420 Diabetes, 440 Kidney transplant, 662 Sickle cell anemia crisis, and 758 Childhood behavioral disorder. Each group has 4 severity levels of illnesses (SOI) and 4 risk levels of mortality (ROM) while the DRG and Medicare Service – Diseases Related Groups (MS-DRG) have only a single severity and risk of mortality per group. For example, there are multiple diagnosis codes for asthma and an encounter might have asthma as principal diagnosis or a secondary diagnosis and if the encounter was primarily for asthma treatment, then the APR-DRG code will be 141 and all asthma encounters will be assigned the same APR-DRG code. In our internal system we code inpatient encounters to APR-DRG as well as DRG. We have available from our PRD back through 2009 [17], including Emergency Room (ER), Ambulatory Surgery (AS), and Observation (OBS) encounters.

Methods

Singular value decomposition

In general, the Singular Value Decomposition method is a method for decomposition of any matrix $A \in R^{MxN}$ where M ≥ N in a product of UV^T , where $U \in R^{Mxk}$ and $V \in R^{Nxk}$ [18,19]. Since any rank *k* matrix can be decomposed in such a way, and any pair of such matrices yields a rank *k* matrix, the problem becomes as an unconstrained minimization over pairs of matrices (U ,V) with the minimization objective

$$\begin{aligned} f(U,V) &= \min\|A - A^{(k)}\|_2^2 \\ &= \min\left(\sum_{n=1}^{N}\sum_{m=1}^{M}|A(m,n) - A^{(k)}(m,n)|^2\right) \qquad (1) \\ &= \min\|A - U^{(k)}V^{(k)^T}\|_2^2 \\ &= \min\left\{\left[A - U^{(k)}V^{(k)^T}\right]^T\left[A - U^{(k)}V^{(k)^T}\right]\right\} \end{aligned}$$

Where $A^{(k)}$ is a rank *k* approximation of matrix A. To find the optimum choices of U,V in l_2 norm sense [20,21], the partial derivatives of the objective *f* (*U,V*) with respect to U,V are

$$\frac{\partial f(U,V)}{\partial U} = 2\left(UV^T - A\right)V \qquad (2)$$

$$\frac{\partial f(U,V)}{\partial V} = 2\left(VU^T - A^T\right)U \qquad (3)$$

Solving $\frac{\partial f(U,V)}{\partial U} = 0$ for U yields $U = AV(V^TV)^{-1}$. By considering an orthogonal solution, then U = Λ is diagonal such that U = AV. Substituting back into $\frac{\partial f(U,V)}{\partial V} = 0$, we have

$$VU^TU - A^TU = V\Lambda - A^TAV = 0 \qquad (4)$$

The columns of V are mapped by A^TA to multiples of themselves, i.e., they are eigenvectors of A^TA. Therefore, the gradient $\frac{\partial f(U,V)}{\partial (U,V)}$ vanishes at an orthogonal (U,V) if and only if the columns of V are eigenvectors of A^TA and the column of U are eigenvectors of AA^T, scaled by the square root of their eigenvalues [18,19]. More generally, the gradient vanishes at any (U,V) if and only if the columns of U are spanned by eigenvector of AA^T and the columns of V are spanned by eigenvector of A^TA. In term of the singular value decomposition, $A = U_oSV_o^T$ the gradient vanishes at (U,V) if and only if there exist matrices $P_U^TP_V = I \in R^{kxk}$ such that $U = U_O\,SP_U$ and $V = V_O\,P_V$. Thus, using singular eigenvectors that corresponds to the largest singular values can represent the global properties (i.e., feature vectors) of A with satisfying the minimization under l_2 norm sense [19].

Low rank matrix decomposition

Suppose that it is desired to represent matrix $X \in R^{MxN}$ as a sum of simple rank one matrices so as to capture the nature of the matrix in which matrix X is to be represented by the summation of r, i.e., rank of matrix. In this case, the outer products can be written as:

$$X = \sum\nolimits_{i=1}^{r} u_i v_i^T \tag{5}$$

Where $X \in R^{M\ x\ N}$, $\{u_1, u_2, \ldots, u_r\}$ and $\{v_1, v_2, \ldots, v_r\}$ vectors each represents linearly independent column vectors with dimensions M and N, respectively. The constituent outer product $u_i v_i^T$ is rank one in which the MxN matrix whose column (row) vectors are each a linear multiple of vector $u_i(v_i)$. To be more precise, a necessary condition is that the vector set $\{u_1, u_2, \ldots, u_r\}$ must form a basis for the column space of matrix X and the vector set $\{v_1^T, v_2^T, \ldots, v_r^T\}$ should form a basis for the row space of matrix X. It is noted, however, that there will exist an infinite number of distinct selections of these basis vectors for the case r ≥ 2. It then follows that there will be an infinite number of distinct ranks when the decomposition of a matrix has rank r ≥ 2. The ultimate selection to be made is typically based on the application as well as computational considerations. To provide a mathematically based method for selecting the required basis vectors, let us consider the functional relationship

$$f_k(\{u_i\},\{v_i\}) = ||X - \sum\nolimits_{i=1}^{k} u_i v_i^T||_p \tag{6}$$

For $1 \le k \le r$ and $p = 1{,}2$ where the integer k ranges in the interval $1 \le k \le r$.

It can be readily shown that the function (6) represents a convex function of the set $\{u_i\}$ for a fixed set of $\{v_i\}$ and vice-versa. For the proof, please refer to [22]. The convexity property is important since it ensures that any local minimum of f_k (v) (i.e., u is fixed) and vice-versa is also a global minimum. With regard to the above equation, a specific selection of the vector sets $\{u_1,u_2,\ldots,u_k\} \in R^M$ and $\{v_1,v_2,\ldots,v_k\} \in R^N$ is to be made so as to minimize this functional. The optimal selection will then provide the best rank k approximation of matrix X in the lp norm sense, as designated by

$$X^{(k)} = \sum\nolimits_{i=1}^{k} u_i^o v_i^{o^T} \tag{7}$$

This optimal matrix is said to be the best rank k approximation of matrix X in the lp norm sense. For convenience, we express equation (5) in a normalized form as:

$$X^{(k)} = \sum\nolimits_{i=1}^{k} \sigma_i^o u_i^o v_i^{o^T} \tag{8}$$

Where $||u_i^o||_p = ||v_i^o||_p = 1$ and σ_i^o are positive scalars. The most employed matrix decomposition procedure is the Singular Value Decomposition (SVD). The SVD method provides an effective method for mitigating the deleterious effects of additive noise and is characterized by the function $f_k(\{u_i\},\{v_i\})$ in the l_2 norm sense, that is

$$||X - X^{(k)}||_2 = \sqrt{\sum\nolimits_{n=1}^{N} \sum\nolimits_{m=1}^{M} \left|X(m,n) - X^{(k)}(m,n)\right|^2} \tag{9}$$

The use of the l_1 norm criterion can be of practical use when analyzing data that contains data outliers. Namely, it would be useful to express this equation (9) as an objective function that optimizes the best rank k approximation of matrix $X \in R^{MxN}$ as measured for the case of the l_1 norm criterion. That is

$$||X - X^{(k)}||_1 = \sum\nolimits_{n=1}^{N} \sum\nolimits_{m=1}^{M} \left|X(m,n) - X^{(k)}(m,n)\right| \tag{10}$$

In order to attempt to find the optimum solution which minimizes the objective function (10), we introduce a concept, called Alternating Optimization, This optimization concept is explained a detailed below.

Alternating optimization

We can rewrite the equation (10) in term of matrices U and V as $f(U,V) = ||X - UV^T||_1$ by fixing U, then objective function becomes:

$$f(V) = ||X - U_{fix} V^T||_1 \tag{11}$$

Where $X = \left[\overrightarrow{x_1}\,\overrightarrow{x_2}.\ .\ .\ .\ .\overrightarrow{x_n}\right], V = \left[\overrightarrow{v_1}\,\overrightarrow{v_2}\,.....\,\overrightarrow{v_k}\right]$ and similarly the column of V^T are denoted by $V^T = [\tilde{v}_1\,\tilde{v}_2.\ .\ .\ .\ .\tilde{v}_n]$. It is straightforward to see that f (V) can be rewritten as a sum of independent criteria

$$f(V) = \sum_{i=l}^{n} \left\|\vec{x}_1 - U_{fix}\tilde{v}_i\right\|_1 \tag{12}$$

where each $||\tilde{x}_i - U_{fix}\tilde{v}_i||_1$ term may be minimized independently by selecting an appropriate. The solution method for each of these subproblems is given in [23]. Grouping the resulting $\tilde{v}_i$ together to obtain V^T, we get a solution for equation (11). On the other hand, by fixing V, the objective function can be expressed as:

$$\begin{aligned} f(U) &= ||X - UV_{fix}^{T}||_1 \\ &= ||X^T - V_{fix}U^T||_1 \end{aligned} \tag{13}$$

And a similar method may be used to solve for U. The iteration process proceeded by finding $\tilde{v}_i$ and then finding $\tilde{u}_i$ (i.e., the alternating optimization) is continued until a stopping criterion is met (i.e., the matrix from two successive iterations are sufficiently close). For example, $||X_{i-1}^{(k)} - X_i^{(k)}||_2 < \varepsilon, \varepsilon = 10^{-7}$. However, it must be noted that finding a global minimum is not guaranteed. In the following section, we establish a guideline for selection of the stopping criteria.

Selection criterion

In this section, let us direct our attention to the selection criteria for the initial choice for *U*, where $U \in R^{Mxk}$. We note that for the following cases where (i) rank $k = 1$ approximation and (ii) rank $1 < k \leq r$, then r = rank (X). In order to take the global data into account, a good choice of initial value of *U* for a rank $k = 1$ (i.e., $U \in R^{Mx1}$) approximation may be obtained as follows. First, we compute the l_1 norm of each column vector in *X*, and denoted this norm by $x_1^c, x_2^c, \ldots, x_n^c$. Next compute the $l_1$norm of each row vector in *X*, and denoted this norm by $x_1^r, x_2^r, \ldots, x_m^r$. Now we find the maximum value in $\{x_1^c, x_2^c, x_n^c, x_1^r, x_2^r, \ldots, x_m^r\}$. If the maximum corresponds to a column norm, say from column j, then chose that column (i.e., $U = X\ (:,j\ )$) as the initial choice for *U*. If the maximum corresponds to a row norm, say row I, then we start with the transposed form of the criterion in (11) and we chose that row (i.e., $V^T = X\ (i,:)$) as the initial choice for V^T. We can also extend the previous concept to find the initial choice for *U* for the rank $k = 2$. Essentially, we apply the rank one approximation twice in succession. Therefore our objective function can be expressed as:

$$\begin{aligned} \min||E_2||_1 &= \underbrace{\min}_{u_1,u_2,v_1,v_2} ||X - [u_1\ u_2][v_1\ v_2]^T||_1 \\ &= \underbrace{min}_{U,V} ||X - UV^T||_1 \end{aligned} \tag{14}$$

Where $U = [u_1 u_2]$ and $V = [v_1 v_2]$, u_1, u_2, v_1, v_2 are vectors. Therefore the initial choice for U (rank $k = 2$) is $U = [u_1 u_2]$ (i.e., two largest l_1 column or row norm). In a similar fashion, a selection criterion for the initial for U for rank k ($1 < k \leq r$) can be also obtained. Thus the column space of *X* (i.e., U) represents a feature vector that is considered as a global property (i.e., the best low rank approximation) of *X* that minimizes the above objective function under l_1 norm sense [22].

Convergence subsequence

The error sequence happened in each iteration can be expressed as:

$$E_i(U,V) = ||X - X_i^{(k)}||_p \text{ where } \mathrm{X}_i{}^{(k)} = U_i V_i^T \text{ and } p = 1,2 \tag{15}$$

Since the error sequence is bounded below (i.e., $E(U,V) \geq 0$)) we have

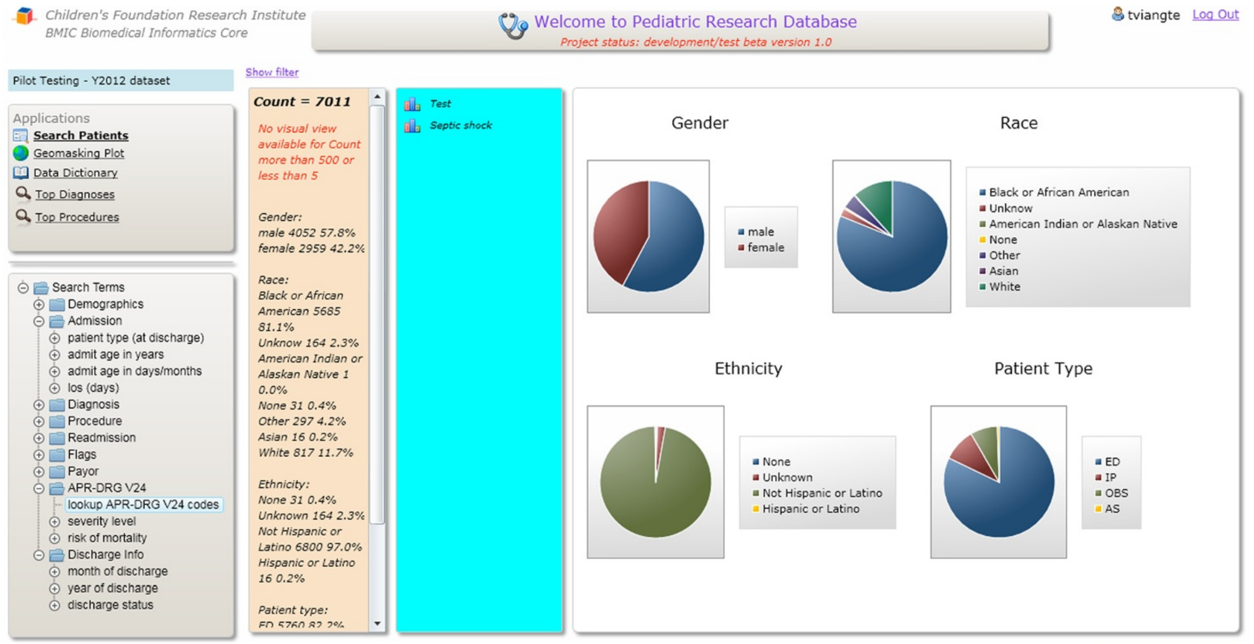
$$E_i(U,V) = ||X - X_i^{(k)}||_p \leq E_1 \text{ where} E_1 \geq 0 \tag{16}$$

And $\lim_{i \to \infty} E_i = E_{final} \geq 0$. Therefore the entire infinite length sequence lies inside a hypersphere (i.e., a closed and bounded set of points) of finite volume centered at *X* and with a radius of E_1. Since this hypersphere has finite volume, it is possible to construct a finite number of smaller hypersphere, each with radius $\varepsilon > 0$, such that the union of all these small hyperspheres contains the large hypersphere of radius E_1. For all $\varepsilon > 0$ there will be at least one hypersphere of radius ε containing an infinite number of points of the sequence. Thus, there is at least one cluster point. The cluster point is the limit of a convergent subsequence. Therefore, we know that the sequence of $X_i^{(\mathrm{k})}$, produced by the algorithm must contain at least one convergent subsequence.

Feature extraction methodology

For the purpose of this preliminary study, we acquired de-identified data sets from PRD that demonstrate patient visits in year 2012. The total number of observations includes 92,175 encounters. Among all encounters, we selected encounters that have APR-DRG code = 141 Asthma, 144 Respiratory signs & minor diagnoses, 131 Cystic fibrosis – pulmonary disease, and 132 BPD & chronic respiratory for our initial datasets. The total number of meeting criteria is 8,895 encounters for 7,011 distinct patient records (see Figure 1 and Figure 2). Among all patients, 57.8% (4,052) were male, 11.7% (817) were white, and 81.1% (5,685) were black or African-American. The PRD has the UTHSC Institutional Review Board (IRB) approval for the creation and maintenance of the database. The waiver applies to the medical records of patients who received care in 2009 or later.

The text parsing software and natural language toolkit [24] (written in Python) were used to parse all encounter data sets for this preliminary study. If $X = [x_{\mathrm{ij}}]$ defines the $m \times n$ term-by-encounter matrix for decomposition. Each element or component x_{ij} of the matrix *X* defines a weighted frequency at which term *i* occurs in encounter *j*,



Count = 7011

No visual view available for Count more than 500 or less than 5

Gender:
male 4052 57.8%
female 2959 42.2%

Race:
Black or African American 5685 81.1%
Unknow 164 2.3%
American Indian or Alaskan Native 1 0.0%
None 31 0.4%
Other 297 4.2%
Asian 16 0.2%
White 817 11.7%

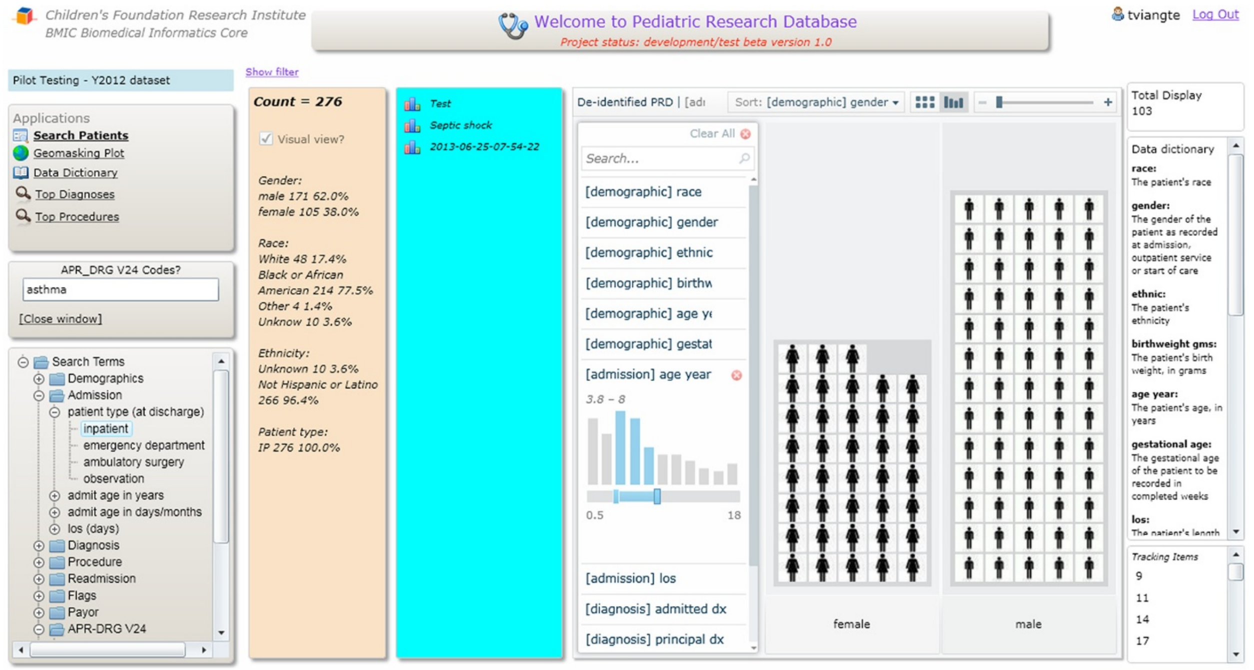
Ethnicity:
None 31 0.4%
Unknown 164 2.3%
Not Hispanic or Latino 6800 97.0%
Hispanic or Latino 16 0.2%

Patient type:
ED 5760 82.2%

Figure 1 **The PRD search cohort demonstrates the ability to search for possible asthma cases based on APR-DRGs.**

where term *i*∈ {gender, age, discharge status, admitting diagnosis, secondary diagnoses, principal diagnosis, principal procedure, secondary procedures}. The corpus stop words from NLTK were used to filter out unimportant terms.

In evaluating the classification performance, we randomly selected subset 1,200 encounters and divided into a number of four subsets of equal size (i.e., four-fold cross validation). The system is trained and tested for four iterations (see Figure 3).



APR_DRG V24 Codes?
asthma

Count = 276

☑ Visual view?

Gender:
male 171 62.0%
female 105 38.0%

Race:
White 48 17.4%
Black or African American 214 77.5%
Other 4 1.4%
Unknow 10 3.6%

Ethnicity:
Unknown 10 3.6%
Not Hispanic or Latino 266 96.4%

Patient type:
IP 276 100.0%

Total Display
103

Data dictionary

race:
The patient's race

gender:
The gender of the patient as recorded at admission, outpatient service or start of care

ethnic:
The patient's ethnicity

birthweight gms:
The patient's birth weight, in grams

age year:
The patient's age, in years

gestational age:
The gestational age of the patient to be recorded in completed weeks

los:
The patient's length

Tracking Items
9
11
14
17

Figure 2 **The PRD visual analytic feature demonstrates the ability to sort and filter by various criteria (tree view sorted by patient's gender and filtered by patient's age).**

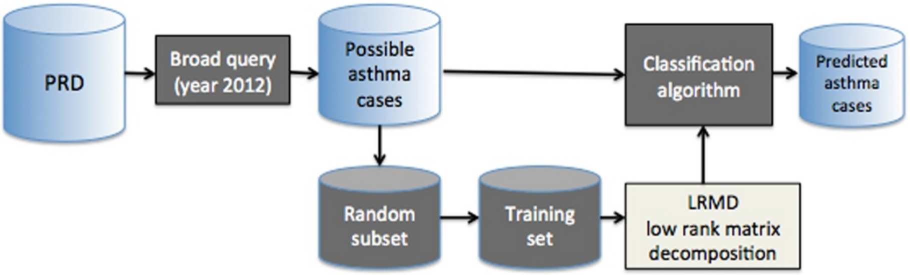


Figure 3 Classification workflow.

In each iteration, three subsets of data are used as training data and the remaining set is used as testing data. In rotation, each subset of data serves as the testing set in exactly one iteration. The rank U used to test the LRMD was $k = 4$. Hence, the U and V matrix factors were number of terms × 4 and 4×1200, respectively. The percentage of possible asthma encounters used for training in our testing was 900 encounters and the remaining 300 encounters were used for testing our classifier. The initial matrix factors U and V were selected to meet our Selection criterion (see Selection Criterion) and alternating iteration was continued until the matrix from two successive iterations are sufficiently close (see Alternating optimization). All classification results were obtained using Python version 2.7.4.

Results

Table 1 demonstrates an example of dominant features for the classifier, when applied to training data sets (randomly selected 900 out of a 1,200 encounters). We note that among all features, admitting diagnosis = 786.07 (wheezing), secondary diagnosis = 786.05 (shortness of breath), age 4–8, and having family history of asthma (ICD-9-CM = v175) would potentially progress toward asthma, i.e., APR-DRG code = 141 asthma. The 2nd Feature shows asthma patients with pneumonia condition (ICD-9-CM = 486.00) during the length of stay in a hospital. The 4th Feature demonstrates the connection between asthma symptoms and another pulmonary condition known as bronchitis symptom (ICD-9-CM = 466.0). When the two conditions co-exist, bronchitis can cause patients with asthma to make their asthma symptoms worse, i.e., an asthma attack.

To evaluate the performance of our classifier for this preliminary study, we plot a receiver operating characteristic (ROC). Figure 4 shows the receiver operating characteristic (ROC) curves (true positive rate versus false positive rate).

Our goal is to maximize the number of true positives (correctly diagnosed asthmas) with an acceptable number of false positives (false alarm). From Table 2 and Figure 4, we note that as sensitivity goes up, a specificity goes down.

Table 1 Example of dominant features using LRMD

Variables	1st Feature	2nd Feature	3rd Feature	4th Feature	5th Feature
admitting diagnoses (ICD-9-CM)	786.07 493.90	786.07 786.09	786.07 493.92	786.07	786.07 493.92
secondary diagnoses (ICD-9-CM)	v175 530.81 786.05	486.00 692.9 v175	786.05 v175 785.0 v174.9	786.05 780.60 692.9 786.2 466.0 v175	v175 692.9 785.0 787.03
principal diagnoses (ICD-9-CM)	493.92 494.90 493.91	493.92 493.90	493.92 786.06	493.92 493.91	493.92
principal procedures (ICD-9-CM)	N/A	939.4	N/A	N/A	939.4
age (year)	4-7	3.5-7	4-6.5	4-6	4.5-8
gender	male	female	female	female	male
discharge status	home	home	home	home	home

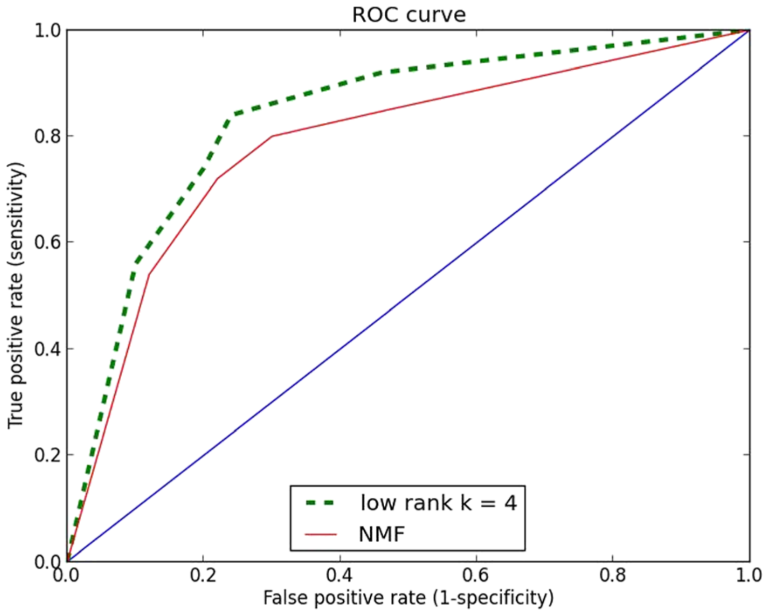


Figure 4 **The receiver operating characteristic (ROC) shows sensitivity and specificity of our classifier.**

We can begin to see the trade-offs when we insist an higher sensitivity (fewer missed asthmas) the result is lower specificity, and more false positive. In practice, it is much worse to miss a asthma than to endure unnecessary treatment, so we tend to choose a higher sensitivity cut off (e.g., cutoff score > 0.65 or > 0.75). As it is apparent from Table 2, LRMD yields very promising results for disease identification and classification. However, we still have much work to do to enhance the LRMD classifier and it is discussed further below.

Discussion

The results presented in this paper should not be taken as an accurate representation of our patient data (as it does not include all the data records). These data are meant to demonstrate the potential of PRD and the feasibility of data mining technique using LRMD. Additional experiments with a larger number of features (rank $k > 4$) and encounter data sets (2009 – 2012) should produce better models to capture the diversity of contexts described by those encounters. Using ICD-9-CM has limitations because they are generally used for billing purposes and not for clinical research. We are planning to access free-text fields in the near future, such as physician and clinician notes, and include them into our classifier. Additional socio-demographic variables such as incomes, type of insurance, environment, nutrition, genome and comorbidity covariants could potentially be added to the model to support the evaluation of potential causes for readmission.

Conclusions

Using data mining technique to learn from past examples within rich data sources such as electronic medical records not only permits users to detect expected events, such as might be predicted by models, but also helps users discover the unexpected patterns and relationships that can then be examined and assessed to develop new insights. We hope that learned rules from the LRMD technique will greatly advance progress toward the goal of identifying high risk of pediatric asthma patient and help support clinical decisions.

Table 2 Sensitivity and specificity

Cutoff score for similarity to features in training set (1 = perfect correlation and 0 = no correlation)	LRMD		NMF	
	Sensitivity	Specificity	Sensitivity	Specificity
> 0.65	0.92	0.54	0.85	0.53
> 0.75	0.84	0.76	0.8	0.7
> 0.85	0.74	0.8	0.72	0.78
> 0.9	0.56	0.9	0.54	0.88

Competing interests

The author declares that they have no competing interests.

Acknowledgements

The author thanks the UTHSC department of ITS Computing Systems and Office of Biomedical Informatics for use of informatics resources and collaborations. We gratefully acknowledge Rae Shell and Grady Wade for assistance on proofreading and providing good comments. This work was supported by the Children's Foundation Research Institute (CFRI).

Received: 20 July 2013 Accepted: 22 August 2013
Published: 28 September 2013

References

1. Chandra Shekar DV, Sesha Srinivas V: *Clinical Data Mining An Approach for Identification of Refractive Errors.* Hong Kong: Proceedings of the International MultiConference of Engineers and Computer Scientists 2008 Vol I IMECS 2008; 2008. 19-21 March.
2. Palaniappan S, Ling C: **Clinical Decision Support Using OLAP With Data Mining.** *IJCSNS International Journal of Computer Science and Network Security* September 2008, **8:**9.
3. Prather JC, *et al*: **Medical data mining: knowledge discovery in a clinical data warehouse.** *Proc AMIA Annu Fall Symp* 1997:101–105.
4. Chae YM, *et al*: **Data mining approach to policy analysis in a health insurance domain.** *Int J Med Inform* 2001, **62**(2-3):103–111.
5. Hedberg SR: **The data gold-rush.** *Byte* 1995, **20**(10):83–88.
6. Mohri M, Rostamizadeh A, Talwalkar A: *Foundations of Machine Learning.* New York: The MIT Press; 2012.
7. Huang Z: **Extensions to the *k*-means algorithm for clustering large data sets with categorical values.** *Data Mining and Knowledge Discovery* 1998, **283:**304.
8. Jain AK, Murty MN, Flynn PJ: *Data Clustering: A Review.* ohio: ACM computing surveys; 1999.
9. Neapolitan RE: *Learning Bayesian Networks.* Illinois: Prentice Hall; 2004.
10. Gelman A: **A Bayesian formulation of exploratory data analysis and goodness-of-fit testing.** *International Statistical Review* 2003, **71**(2):369–382.
11. Tom M: *Machine Learning.* McGraw-Hill; 1997:55–58.
12. Grzymala-Busse JW: **Selected algorithms of machine learning from examples.** *Fundamenta Informaticae* 1993, **18:**193–207.
13. Liu WX, *et al*: **Nonnegative matrix factorization and its applications in pattern recognition.** *Chinese Science Bulletin* 2006, **51**(1):7–18.
14. Cemgil AT: **Bayesian inference for nonnegative matrix factorisation models.** *Comput Intell Neurosci* 2009:785152.
15. Berry MW, Gillis N, Glineur F: *Document Classification Using Nonnegative Matrix Factorization and Underapproximation.* IEEE; 2009.
16. Sedman AB, Bahl V, Bunting E, Bandy K, Jones S, Nasr SZ, Schulz K, Campbell DA: **Clinical redesign using all patient refined diagnosis related groups.** *Pediatrics* 2004, **114**(4):965–969.
17. Viangteeravat T: **Giving Raw Data a Chance to Talk: A demonstration of de-identified Pediatric Research Database and exploratory analysis techniques for possible cohort discovery and identifiable high risk factors for readmission.** *Proceeding of 12TH Annual UT-ORNL-KBRIN Bioinformatics Summit* 2013.
18. Srebro N, Jaakkola T: *Weighted Low Rank Approximation.* Washington DC: Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003); 2003.
19. Young E: *Singular Value Decomposition.* http://www.schonemann.de//svd.htm.
20. Cadzow JA: **Signal enhancement: a useful signal processing tool Spectrum Estimation and Modeling.** *Fourth Annual ASSP Workshop* 1988, **162:**167.
21. Cadzow JA: **Minimum l(1), l(2), and l(infinity) norm approximate solutions to an overdetermined system of linear equations.** *Digital Signal Processing* 2002, **12**(4):524–560.
22. Viangteeravat T: *Discrete Approximation using L1 norm Techniques.* Master Thesis: Electrical Engineering, Vanderbilt University; 2000.
23. Cadzow JA: *Application of the l_1 norm in Signal Processing". Department of Electrical Engineering.* Nashville: Vanderbilt University; 1999.
24. Perkins J: *Python Text Processing with NLTK 2.0 Cookbook.* Birmingham: Packt Publishing; 2010.

doi:10.1186/2043-9113-3-16
Cite this article as: Viangteeravat: **Potential identification of pediatric asthma patients within pediatric research database using low rank matrix decomposition.** *Journal of Clinical Bioinformatics* 2013 **3**:16.