Proceedings of 31st International Society Conference on
Computers And Their Applications CATA 2016, pp. 251-256, April 05, 2016

Principal Component Analysis as an Integral Part of Data Mining in Health Informatics

Chaman Lal Sabharwal[1] and Bushra Anjum[2]
[1]Missouri University of Science and Technology, Rolla, MO-63128, USA
chaman@mst.edu
[2]Amazon Inc, 1194 Pacific St., San Luis Obispo, CA-93401, USA
banjum@amazon.com

Abstract

Linear and logistic regression are well-known data mining techniques, however, their ability to deal with inter-dependent variables is limited. Principal component analysis (PCA) is a prevalent data reduction tool that both transforms the data orthogonally and reduces its dimensionality. In this paper we explore an adaptive hybrid approach where PCA can be used in conjunction with logistic regression to yield models which have both a better fit and a reduced set of factors than those produced by just the regression analysis. We will use example dataset from HealthData.gov to demonstrate the simplicity, applicability and usability of our approach.

keywords: Principal component analysis, Regression analysis, healthcare analytics, big data analytics

1. Introduction

From doctors' fees and medical tests to the price of medications and the cost of hospital stays, health-care costs around the world are skyrocketing. Much of this is attributed to wasteful spending on such things as ineffective drugs, futile procedures and redundant paperwork, as well as missed disease-prevention opportunities. This calls for efficient data reduction mechanisms and diagnostic tools that can help identify the problems and the factors which are most responsible for them. This is where Big Data analysis comes in the picture.

It is estimated that developing and using prediction models in the health-care industry could save billions by using big-data health analytics to mine the treasure trove of information in electronic health records, insurance claims, prescription orders, clinical studies, government reports, and laboratory results. Improving the care of chronic diseases, uncovering the clinical effectiveness of treatments, and reducing readmissions are expected to be top priority use cases for Big Data in healthcare [1]. According to the Harvard School of Public Health publication entitled The Promise of Big Data, petabytes of raw information could provide clues for everything from preventing tuberculosis to shrinking health care costs [6]. Some of the initiatives taken in this domain are discussed below.

Today, there is a significant opportunity to improve the efficiencies in the healthcare industry by using an evidence-based learning model, which can in turn be powered by Big Data analytics [8]. A few examples are provided below. The company Asthmapolis has created a global positioning system (GPS) enabled tracker that monitors inhaler usage by patients, eventually leading to more effective treatment of asthma [9]. Center for Disease Control and Prevention (CDC) is using Big Data analytics to combat influenza. Every week, the CDC receives over 700,000 flu reports including the details on the sickness, what treatment was given, and whether not the treatment was successful. The CDC has made this information available to the general public called FluView, an application that organizes and sifts through this extremely large amount of data to create a clearer picture for doctors of how the disease is spreading across the nation in near real-time [3]. Another area of interest is the surveillance of adverse drug reactions (ADRs), which has been a leading cause of *death* in the United States [4]. It is estimated that approximately 2 million patients in USA are affected by ADRs and the researchers in [11,13] propose an analytical framework for extracting patient–reported adverse drug events from online patient forums such as DailyStrength and PatientsLikeMe.

Simplistically speaking, in all the above examples the researchers are trying to model and predict a dependent phenomenon based on a number of predictors that have been observed. The dependent parameter can be discrete or nominal or even binary/logical. There are two problems at

hand. *First* problem is analyzing whether some event occurred or not given the success or failure, acceptance or rejection, presence or absence of observed simulators. If such a dependence and correlation can be established, then there is a *second* equally interesting problem of optimization. The optimization problem is how we can minimize the set of predictors while still maintaining high prediction accuracy for the dependent variable.

There are several approaches to modeling prediction variables, such as, linear regression analysis, logistic regression analysis, and principal component analysis. Each has its own advantages and disadvantages. A mathematical background to these is presented in section 2. Though regression analysis has been well known as a statistical analysis technique, the understanding of PCA, however, has been lacking in the past by non-academic clinicians. In this paper we explore an adaptive hybrid approach where PCA can be used in conjunction with logistic regression and unsupervised learning to yield models which have both a better fit and reduced number of variables than the models predicted by standalone logistic regression or unsupervised learning. We will apply our findings to a dataset obtained from HealthData.gov that lists the quality ratings of California's hospitals based on their location, medical procedure, mortality rate, no. of cases etc. [7]

The paper is organized as follows. Section 2 presents the mathematical background on linear regression, logistic regression, and PCA. Section 3 describes the hybrid algorithms for regression using PCA. In section 4 we use the hospital rating dataset from HeathData.gov to present experimental results of our algorithm and Section 5 concludes the paper.

2. Background

2.1. Mathematical Notation

Here we describe the mathematical notation for terms used in this paper. A vector is a sequence of variables. All vectors are *column* vectors and are in lower case *bold* letters such as $\mathbf{x}$. The n-tuple $[x_1,..., x_n]$ denotes a *row* vector with *n* elements in lowercase. A superscript T is used to denote the *transpose* of a vector $\mathbf{x}$, so that $\mathbf{x^T}$ is a row vector whereas $\mathbf{x} = [x_1,..., x_n]^T$ is a column vector. This notation is overloaded at some places where the ordered pair $[x_1, x_2]$ may be used as a row *vector*, a *point* in the plane or a closed *interval* on the real line. The *matrices* are denoted with uppercase letters, e.g. A, B. The n**x**n *identity* matrix is denoted by I_n , or simply I when the dimension is implicit in the context. The *elements* $I_{ij} = 0$ if $i \neq j$ and $I_{ii} = 1$ for $1 \leq i, j \leq n$. For vectors $\mathbf{x}$, $\mathbf{y}$, the covariance is denoted by cov($\mathbf{x}$, $\mathbf{y}$), whereas cov($\mathbf{x}$) is used for cov($\mathbf{x}$, $\mathbf{x}$) as a shortcut [2].

If we have *m* vector values $\mathbf{x}_1,..., \mathbf{x}_m$ of an *n*-dimensional vector $\mathbf{x} = [x_1,..., xn]^T$, these m observations are represented by an m**x**n data matrix **A.** The k^{th} row of A is the row vector $\mathbf{x}_k^T$. Thus the (i, j) element of **A** becomes to the j^{th} element of the i^{th} row/observation, $\mathbf{x}_i^T$.

There are several ways to represent data so that implicit/hidden information becomes transparent. For linear representation of vector data, a vector space is equipped with a basis of linearly independent vectors. Usually in data mining, the data is represented as a matrix of row vectors or data instances. Two of the methods for efficient representation of data are regression and principal component analysis, Figure 1.

2.2. Linear Regression

The linear regression for one dependent and *m* independent variables is given as $\boldsymbol{y = b_0 + \sum_{k=1,m} b_k x_k}$ where the error between the *observed* value y_i and *estimated* value $\boldsymbol{b_0 + \sum_{k=1,m} b_k x_{ik}}$ is minimum. For *m* points data, we compute b_k by using the method of least squares that minimizes

$$\sum_{\boldsymbol{i=1,n}} (\boldsymbol{y_i} - b_0 - \sum_{k=1,m} b_k x_{ik})^2$$

Thus linear regression determines a hyper plane which is a least square approximation of the data points. If data is mean centered, then b_0=0. If m=1, it is a regression line. It is always better to mean center the data as it simplifies the computations.

It is important to note that the data points *may not* be at the least distance from the regression line. We show this in the next section, that there is a better least distance line, see Figure 2. For direction vectors and approximation error of data points from the line, see Table 1.

2.3. Principle Component Analysis

In linear regression, the distance between observed point (x_i,y_i) from the line $y=a+bx$, along the y direction is minimized. In principal component analysis, the distance of observed (x_i,y_i) along a direction orthogonal to the line $y=a+bx$, is minimized.

The principal component analysis is a well-known data reduction tool in academia for over 100 years. PCA is beneficial for representing physical world data/objects more clearly in terms of independent, uncorrelated, orthogonal parameters. In the next section we explore an adaptive hybrid approach where PCA can be used not only for data reduction but also to yield models which have a better fit than those produced by using logistic regression or unsupervised learning alone.

PCA and Singular Value Decomposition (SVD) are interchangeably used in the literature. However, there is a

clear distinction between them as can be noted from the discussion below.

Definition 1. For a real square matrix A, if there is a real number λ and a *non-zero* vector **x** such that $A\,\mathbf{x} = \lambda\,\mathbf{x}$, then λ is called an eigenvalue and **x** is called an eigenvector.

Definition 2. For a real matrix A (square or rectangular), if there a *non-negative* real number σ and a non-zero vectors **x** and **y** such that $A^T\,\mathbf{x} = \sigma\,\mathbf{y}$, and $A\,\mathbf{y} = \sigma\,\mathbf{x}$, then σ is called a singular value and **x** and **y** represent a pair of singular vectors. [10]

Note 1. λ can be negative or positive, but σ is never negative.

Note 2. σ^2 is an eigenvalue of covariance matrices AA^T and A^TA.

In data mining, the *m* observations data is represented as an mxn matrix A where each observation is a vector with *n* components. The direction vectors from which there is least deviation, or along which there is most variation, are computed as follows. If A is a real square *symmetric* matrix, then eigenvalues are real and eigenvectors are orthogonal [11]. If we form U as matrix of eigenvectors and D as diagonal matrix of corresponding eigenvalues then A can be written as:

$$A\,U = U\,D \quad \text{or} \quad A = U\,D\,U^T$$

It is called the principal component decomposition (PCA) of A. Since the columns of U are orthogonal eigenvectors of A, then U is orthogonal matrix and its inverse is the transpose, e.g., $U^T = U^{-1}$. Also PCA orders the eigenvalues in the descending order of magnitude. The columns of U and diagonal entries of D are correspondingly arranged. Since the eigenvalues can be negative, the diagonal entries of D are ordered based on absolute values of eigenvalues.

Note 3. For eigenvectors **u** of A, any non-zero multiple of **u** is also an eigenvector. In U, the eigenvectors are normalized to unity. If **u** is a unit eigenvector, then **-u** is also unit eigenvector. The sign can be arbitrarily chosen, some authors make the first nonzero element of the vector to be positive to make them unique. This is really a cool way to make it unique. We do not follow this convention. Since the eigenvectors are ordered, we make the k^{th} element of vector $\mathbf{u}_k$ as positive; if k^{th} element is zero, only then first non-zero element is made positive. This is a better representation of eigenvectors as it is both visually appealing and *gets rid of the asymmetrical ordering in favor of a right handed system of representation*, see Figure 1. Using Matlab svd(A) on a simple set of only two data points, the algorithm generates two orthogonal vectors v_1, v_2 as in Figure 1(a, b). Our algorithm finds green vectors in Figure 1(c), which is a more natural representation with a right handed orientation. The vectors in U and V are such that the data has the most variation along these directions in descending order. Data is most spread along v_1, see in Figure 1 (c).

Next we claim, and present an example, that PCA can give us a better approximation – a better least distance line – than standard linear regression. We have a data set of randomly created 20 points and we use both methods, i.e., linear regression (also termed as least square approximation) and PCA, in Matlab to compute vectors such that the variation of the observed points from the computed vector is minimum. The results are given in Figure 2, where the linear regression is given as the red line and the PCA approximation is given as the blue line.

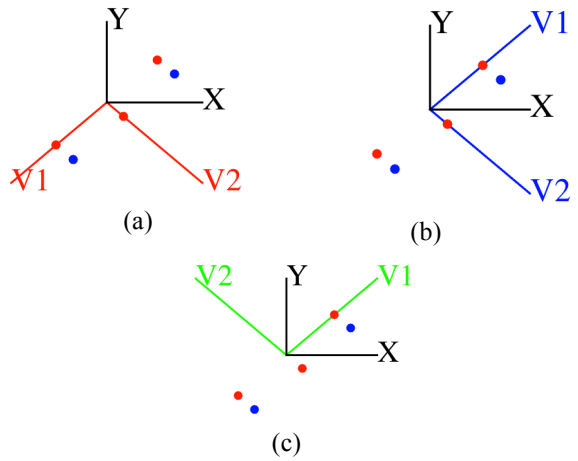

Figure 1: Black set of axes are the standard xy-system. Blue dots are a set of two data points. Red dots are projections of Blue dots on the principal components. The axes v1, v2 in red are the eigenvectors that are computed by Matlab using the simple set of two points, Figure 1(a). In Figure 1 (b) Blue axes v1, v2 are the directions so that each eigenvector is unique making the first non-zero component positive (standard scheme used in the literature). In Figure 1(c), Green v1, v2 are the eigenvectors chosen by our scheme.

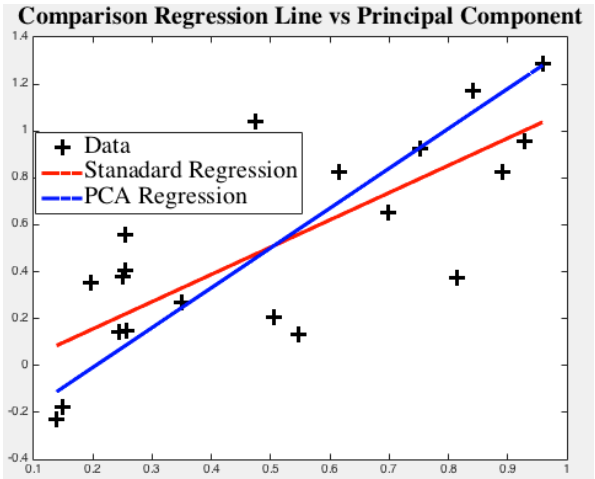

Figure 2: Traditional linear regression line (LSA) is shown in red and a principal component analysis (PCA) line is shown in blue. We can see that the points closer to the blue line are at a lesser distance than the points closer to the red line.

The approximation error, as given in Table 1, indicates that the principal component analysis (PCA) adapted regression line is a better fit than the LSA approximation. Data reduction is attributed to the non-zero eigenvalues of the data matrix A. Since mxn data

matrix is decomposed into A = USV^T where U is mxm, S is mxn, and V is nxn. If there are only k non-zero eigenvalues where $k \leq min(m, n)$, the matrix A has lossless representation by using only k columns of U, k columns of V and kxk diagonal matrix S. If some eigenvalue is very small as compared to others, then it can lead to further data reduction while retaining almost the same information.

Table 1 Comparison of Least Square Approximation

For normal regression line	
Direction vector	[0.642388, 0.766380]
Relative Regression Error	0.391169
For PCA adapted regression line	
Direction vector	[0.514757,0.857336]
Relative Regression Error	0.173438

2.4. Logistic Regression

Linear regression is suitable for data that exhibits linear relation, but this is not always the case. The Logistic model focuses on probabilistic estimates and is applicable to "S-shaped" data [5]. This model is particularly suitable for population growth with limiting condition. Linear and logistic regression (log linear) have been the popular mining techniques due to their simplicity, but their ability to deal with inter dependent data variables is limited. However, this limitation can be removed when they are coupled with PCA.

Example. In this example we have a training data set [12] of 20 students who studied for an exam for given number of hours (horizontal axis) and have passed or failed (vertical axis) the test. Here we determine the trained logistic regression predictor for chances of passing the exam. Fail and Pass are coded numerically as 0 and 1, see Figure 3.

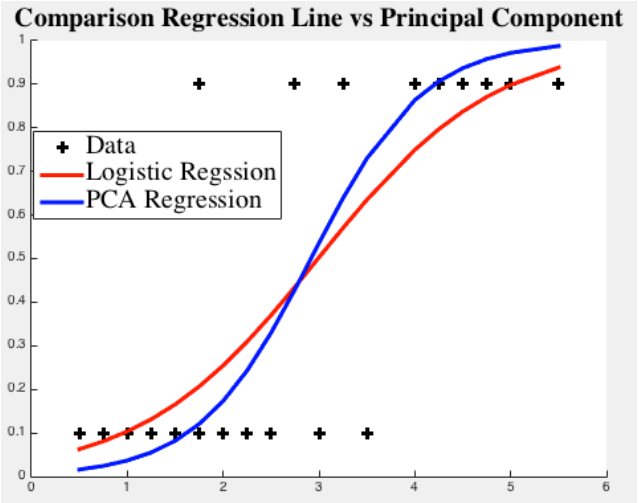

Figure 3: Black dots are data points. Red curve the logistic regression curve computed with standard Logistic regression, blue curve is the one created with hybrid model.

As before, the blue curve (which corresponds to PCA regression curve) shows a much better approximation with 60% less approximation error.

The errors for both the methods are given in Table 2 below.

Table 2. Comparison of Logistic Regression Methods

Comparison of Logistic Regression Methods	
Logistic Regression Relative Error	0.443061
PCA Relative Error	0.157216

2.5 How do we measure the goodness of a model?

In data mining there are standard measures, called gold standard, for labeling and measuring the prediction accuracy. Following metrics are extremely useful in comparing the results of a classification, Precision (P), Recall (R) and F-metrics. They are given as:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

The measure F_1 is Harmonic average of P and R. It is given by

$$F_1 = \frac{2PR}{P + R} \; or \; F_1 = \frac{2TP}{2TP + FP + FN}$$

The reader may consult [12] for further details on these metrics**.**

3. Hybrid Model

Regression analysis is normally designed independent of PCA. PCA is applied to the data, not to the regression algorithm for linear or non-linear logistic regression. In conjunction with PCA, we propose several improvements to the existing practices in designing linear and logistic regression. There are two ways regression is improved: data reduction and hybrid algorithm. As a first step, PCA is used for data reduction. At the second step, standard regression analysis may be used for prediction. This makes PCA and Regression analysis as two separate algorithms in sequence. Further, these steps may be followed by human reviewer before final inference is made. The following paragraph explains why the human reviewer may be necessary in some applications.

Consumers and physicians use different terminologies for the same concept, disease or medicine. In fact, consumers have difficulty in understanding the medical terminology. As a result of misinterpretation, it can be harmful for the consumer. Consumer Health Vocabulary (CHV) is an attempt to bridge the gap, see [11]. Work

continues on extending CHV, which is not complete. Consumer content is available through various social networks where they talk about drugs and their effects, adverse reactions. Using NLP parsing techniques, the consumer data is parsed in the form of medical expressions and stored in the form of term-frequency inverse-document-frequency (TF-IDF). The data is bound to have redundancies, which can be eliminated using principal of component analysis. The CHV is a map from consumer expressions to unified medical language system (UMLS). For example, the terms "head ache" and "cranial pain" are both synonyms of the UMLS concept "headache." If the consumer expression is not in CHV, then it can be a candidate to extend CHV. The English language is ambiguous. The NLP does not check for semantics, for example if a consumer says "heart attack", it will be interpreted as it says. For a young person, it may mean hardship, while for an old person it may mean heart disease. This is where the human reviewer comes in. Before the expression can be a candidate for confirming CHV term, the human reviewer validates or invalidates the consumer expression as medical term against the context where the expression is used, recall TF-IDF structures.

As noted in Figure 2, the error in regression line approximation is much larger than with the line created with PCA. For improved performance in the second step, we create a hybrid logistic regression coupled with PCA adapted regression line. The additional advantage occurs when we make use of this knowledge in designing better logistic regression for prediction in mortality analysis, see Figure 3. This greatly enhances the existing algorithms. Here is the hybrid algorithm in 2D, it can be easily extended to higher dimensions. The following two algorithms replace the existing traditional algorithms.

Improved linear Regression

Input: array of data points (*x, y*)

Output: line *y=a + bx*

Method:

Traditional: compute a and b, by minimizing

$$\sum_{i=1,n} (y_i - a - bx_i)^2$$

Let error1 be the computed minimum error value.

New: compute a and b, by minimizing

$$\sum_{i=1,n} \left(\frac{y_i - a - bx_i}{\sqrt{(1+b^2)}}\right)^2$$

This step is easily adapted by using covariance matrices, as done in principal component analysis. Let error2 be the minimum error value.

Compare error1 and error2.

From Table 1, it shows that error2 is much smaller than error1. Thus adaptive PCA approach regression line is better.

Improved Non-linear Logistic Regression

Input: array of data points (*x, y*)

Output: non-linear PCA adapted logistic function

Method: For logistic regression, map

$$y \rightarrow log_e\left(\frac{y}{1-y}\right)$$

Apply *improved* regression line to *y* values computed from (*new approach*) line *y=a + bx*

Map y values back

$$y \rightarrow \frac{e^y}{1+e^y}$$

This gives the hybrid logistic regression function, see Figure 3 and for approximation error see Table 2. We have designed the improved regression line and logistic regression algorithms by adapting principal component analysis.

4. Experimental Analysis

We used California Hospital Rating dataset from HeathData.gov [7] for our experiments. Our experiments are focused on both data reduction using PCA and hospital rating prediction using improved Logistic Regression. The data is published by the California's Office of Statewide Health Planning and Development (OSHPD). It consists of information on 11,169 patients in the year 2012-2013, ranging over 55 counties in California, 17 types of diseases (labeled as causes of mortality) and the rating -- acceptable or unacceptable – of the hospital. The hospitals were classified on county, general description, OSHPD ID, procedure/condition, risk adjusted mortality rate, number of deaths, number of reported cases per procedure/condition and the hospital location. There were two types of ratings for each hospital for treatment recommendation: unacceptable (worse), acceptable (as expected, better).

For classification, this is a fair size of data for Rating Hospital for treatment, about 12000 cases. The goal is not data mining per se, but to show the feasibility of improved algorithms over the existing algorithms and data reduction to rating hospitals. The reduction in one attribute reduces the data size by 9%.

After updating the data set by ignoring incomplete records with missing information, the data set was reduced to 3033 records with six attributes (County, OSHPD, Procedure/Condition, Risk Adjusted Mortality Rate, # of Deaths, # of Cases, Hospital Rating). We took the last attribute, i.e., Hospital Rating as our dependent variable on

which classification would be performed. We applied PCA on the data with first five attributes.

For experiment, we created two versions of the data set: First version, raw data and second version, mean centered with unit standard deviation. We calculated the eigenvalues next. Four of the five eigenvalues were zero (upto at least ten digits after decimal), the non-zero eigenvalue is shown in Table 3. This indicated that only one new attribute was sufficient to rate the hospitals. But this does not specify which original attributes contributed to the reduction. The principal components on normalized data are more realistic in this case as the normalized data attribute values are evenly distributed. For nominal attributes, mapping nominal to numerical can make a difference. However covariance and correlation approaches are complementary.

The principal component corresponding to the non-zero eigenvalue is a linear combination of original attributes. Each coefficient in it is a contribution of the original data attributes, i.e., each coefficient is a fractional contribution of the original data attributes. It is clear that the two (#deaths and # cases) of the five coefficients are more dominant than the others, however raw data analysis found only one dominant coefficient. In either case, the contribution of the original two attributes is more than 95%. After eliminating the other attributes, we compute the approximation error due to dimension reduction (from five to two attributes). As can be seen in the Table 3, the error is less than is 0.05 percent.

Table 3. PCA Eigen pairs and Error in Data Reduction.

	Raw Data	Normalized Data
Eigen Values	-71.5188	-108.94
Eigen Vectors	$\begin{Bmatrix} 0.0601 \\ -0.1819 \\ 0.1183 \\ 0.7228 \\ 0.6534 \end{Bmatrix}$	$\begin{Bmatrix} 0.0856 \\ 0.0264 \\ 0.0166 \\ 0.0461 \\ 0.9948 \end{Bmatrix}$
Errors	0.0542	0.0347

So the PCA analysis shows that even after 60% reduction, using only 40% of data, the precision is almost the same whereas the gain in computation performance is significant. For recall, the reduced data regression misses more negatives see Table 4. It is preferable to miss less positive than more negatives. Table 4 corresponds to traditional logistic regression Table 5 corresponds to hybrid logistic regression algorithm. It shows that hybrid algorithm consistently out performs the traditional algorithms.

Table 4. Error Comparison Metrics *Traditional* Algorithm

	Raw	PCA 40% Data
Precision	0.952	0.951
Recall	0.807	0.744

Table 5. Error Comparison Metrics *Hybrid* Algorithm

	Raw	PCA 40% Data
Precision	0.965	0.957
Recall	0.897	0.759

5. Conclusion

We presented a hybrid algorithm that adaptively uses PCA to improve the linear and logistic regression algorithms. With experiments we have shown the effectiveness of the enhancement. All data mining applications that dwell on these two algorithms will benefit extensively from our enhanced algorithm but they are especially useful in healthcare informatics where regression is the most popular statistical analysis tool. We applied our algorithms to the quality ratings dataset for California Hospitals to demonstrate its simplicity and applicability.

6. References

[1] Allen Bernard. Healthcare Industry Sees Big Data As More Than a Bandage, *CIO*, August 5, 2013.

[2] Christopher M. Bishop. *Pattern Recognition and Machine Learning*, Springer, 2006.

[3] Julie Bort. How the CDC Is Using Big Data To Save You From The_Flu, *Business Insider*, December 13, 2012.

[4] Brant Chee, Richard Berlin and Bruce Schatz. Predicting Adverse Drug Events from Personal Health Messages. *In AMIA Annual Symposium Proceedings,* pages. 217–226, 2011.

[5] Anupam Goel, Richard G Pinckney and Benjamin Littenberg. APACHE II Predicts Long-term Survival in COPD Patients Admitted to a General Medical Ward. *Journal of General Internal Medicine*, pages 824-830, October 2003.

[6] The Promise of Big Data. *Harward School of Public Health Magazine*, pages 15-43, 2012.

[7] HealthData.gov dataset California Hospital Ratings http://www.healthdata.gov/dataset/california-hospital-inpatient-mortality-rates-and-quality-ratings-2012-2013.

[8] The Global Use of Medicines: Outlook Through 2016. *IMS Institute Reports*. IMS Institute for Healthcare Informatics, pages 1-36, 2012.

[9] Shel Israel, Contextual Health vs The Elephant in the Hospital. *Forbes Tech*, May 23, 2013.

[10] Jim Hefferon. *Linear Algebra*. http://joshua.smcvt.edu/linearalgebra, 2014.

[11] Ling Jiang, Chris Yang and Jiexun Li. Discovering Consumer Health Expressions from Consumer-Contributed Content. *Social Computing, Behavioral-Cultural Modeling and Prediction*, Lecture Notes in Computer Science Vol. 7812 p.164-174, 2013.

[12] Wikipedia. https://en.wikipedia.org/wiki/F1_score, August 2015.
[13] Christopher Yang, Ling Jiang, Haodong Yang, Mi Zhang. Social Media Mining for Drug Safety Signal Detection. *In Proceedings of the 2012 international workshop on Smart health and wellbeing*, pages 33-40, Maui, Hawaii, USA, October 29, 2012.