



Research Journal of Pharmaceutical, Biological and Chemical Sciences

Big Data: A dimensionality Reduction and Attribute Selection using PCA for Diabetic Data bases.

S Santhosh Kumar*.

Department of Computer Science, Government College for Women (A), Kumbakonam, Tamil Nadu, India.

ABSTRACT

Noninsulin dependent diabetes mellitus (NIDDM) is type 2 diabetes that is caused due to high blood sugar. The type 2 diabetes is extremely common. The clinical data of diabetic data bank contains large collection of data sets with complex data values known as big data. The big data banks are growing with the addition of new information with existing one makes more complex to predict necessary information. The important complexity of diabetic data set is attribute selection. The attribution selection and reduction is one of the important issues in data acquisition. Large and multidimensional data sets are having many attributes with related and unrelated values. The improper selection of attributes gives inconsistent results which leads poor performance. This paper focuses development of better information retrieval mechanism to search patient records based on attribute selection.

Keywords: Attribute selection, dimensionality reduction, PCA, Diabetic data set,

**Corresponding author*

INTRODUCTION

Clinical data management is a very large field which has clinical trial data gathered both manually and automatically. These clinical data sets are used in wide area of applications. The remarkable applications are disease diagnosis, drug development, drug design, patient record keeping, drug prescription etc. In technological point of view clinical data management comes under a computational biology in data mining which requires both biological knowledge and computational knowledge respectively. Especially medical data set requires consistent support and use of data mining techniques. The important point is the priority must be given to data only. Based on the generic nature of data can only decides what kind of technique is going to be used. The diabetic data bases are more complex to store, search and analyse the data. When compare with other diseases it is a quiet common disease and does not have any specialised criteria such as age, gender, locality, and symptoms etc., diabetic data sets are always holding multiple attributes with mixed values. The present work focuses the use of data mining technique for the meaningful extraction of patient record from large data bases.

In the case of medical data for example the disease diagnosis of a patient can be described by gender, symptoms, test type etc., it gives better results when the data set is searched with limited or selective attributes. For very large data sets with multiple attributes, it is lack to achieve reliability. In order to solve this issue; the primary or highly weighted attributes must be selected as basic class labels and by that way the dimensionality reduction can be achieved.

About Data Set

Figure 1: Scatter plot of diabetic data set

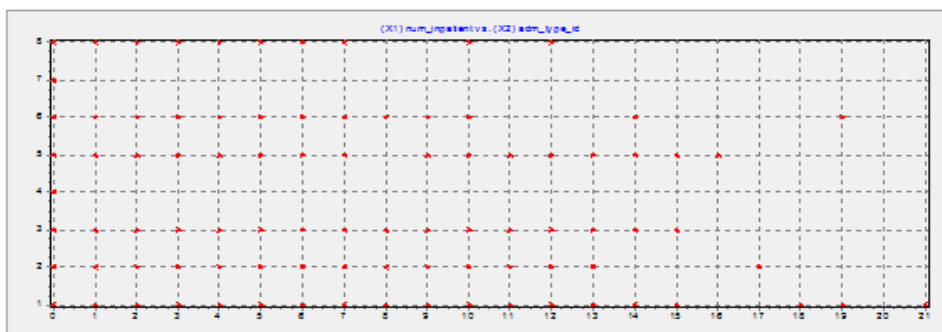


Table 1: Significantly selected Attributes from diabetic data set

Attribute	Category	Informations
Race	Discrete	6 values
gender	Discrete	3 values
age	Discrete	10 values
adm_type_id	Continue	-
dischrg_disp_id	Continue	-
adm_srce_id	Continue	-
tme_in_hsptl	Continue	-
num_emergency	Continue	-
num_inpatient	Continue	-
numbr_diag	Continue	-
change	Discrete	2 values
diabetesMed	Discrete	2 values
readmitted	Discrete	3 values

The data set taken for this work consists of 101767 clinical records with 50 attributes. Each attribute has single or multiple descriptions. This data set is obtained from world wide data management organization. Figure 1 show the diabetic data set representation

For example admission attribute has emergency, urgent, elective, new born, not available, trauma, center, and not mapped descriptions. Similarly all attributes contains atleast two descriptions. This work is focused to predict the categorical values of patients so the attributes which are related to drug and dosage levels are not considered. The primary attributes of patient indicators are taken. The attributes includes race, gender, age, admission details, Diagnosis, diabetes medications. The selected attributes are unrelated parameter with primary classification of discrete and continuo's attributes.

Due to multi dimensionality nature of these attributes it is too complex to select the appropriate attribute. The lack of improper selection of attribute leads to poor results. Inorder to get optimum results, the primary attributes called primary component must be selected before searching patient records. The Principal component analysis (PCA) [3] is an, non-parametric method of extracting relevant information from confusing data sets. In this paper PCA is used as filter technique to reduce dimensionality of an attribute and determine the highest valued attribute relate to required information.

Principal Component Analysis (PCA)

PCA was introduced by Pearson in 1901[1] and Hotelling in 1933[2] to di scribe the variation in a set of multivariate data in terms of a set of uncorrelated variables. The main purposes of a principal component analysis are the analysis of data to identify patterns and finding patterns to reduce the dimensions of the dataset with minimal loss of information. It searches for k n -dimensional orthogonal vectors [5] are represented as $K \leq n$ which combines relative attributes into smaller sets with reduced dimensionality. Each set of attributes are as unit vectors in a direction perpendicular to one another. The highest order of strength of significant vectors are known to be principal components. The least significant vectors are not considered for attribute selection rather it could be considered with significant vectors. The data matrix of n observations on p correlated variables x_1, x_2, \dots, x_p and the transformation of the x_i into p new variables y_i that are uncorrelated. Basically PCA [4] is based on eigenvalue decomposition developed by Marcus and Minc 1988[6]. Eigenvectors are fundamental to principal components analysis which is commonly used for dimensionality reduction in other machine learning applications.

RESULTS AND DISCUSSION

PCA is a measure of the data variance by each of the new coordinate axis. They are used to reduce the dimension of large data sets by selecting only a few modes with significant eigen values and to find new variables that are uncorrelated. It defines A (non-zero) vector v of dimension N is an eigen vector of a square ($N \times N$) matrix A if and only if it satisfies the linear equation.

$$Av = \lambda v$$

where λ is known as the eigen value associated with the eigenvector v . With the use of eigen values the most significant attributes in the diabetic data set are predicted by each attributes mean and standard deviation. The table 2 shows the eigen value of selected attributes with respect to its matrix traces.

Histogram of given attributes in a decreasing order shows disjoint in subsets. The percentage explained and cumulated ranges an expansion and closeness of attributes in set. The partitions results data distribution of diabetic data set with dimensionality reduction. The fig 1 shows the scatter plot of with 3 – dimensional attribute categorisation. The number of diagnosis attribute is first primary data component similarly admission source id and race are secondary attributes respectively.

Table 2: Eigen values

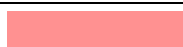




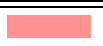
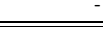
Axis	Eigen value	% explained	Histogram	% cumulated
1	1.421800	20.31%		20.31%
2	1.253523	17.91%		38.22%
3	1.204477	17.21%		55.43%
4	0.946506	13.52%		68.95%
5	0.765587	10.94%		79.88%
6	0.722837	10.33%		90.21%
7	0.685271	9.79%		100.00%
Tot.	7.000000	-	-	-

Figure 1: Scatter plot for primary significant attribute number of diagnosis with it supports attributes

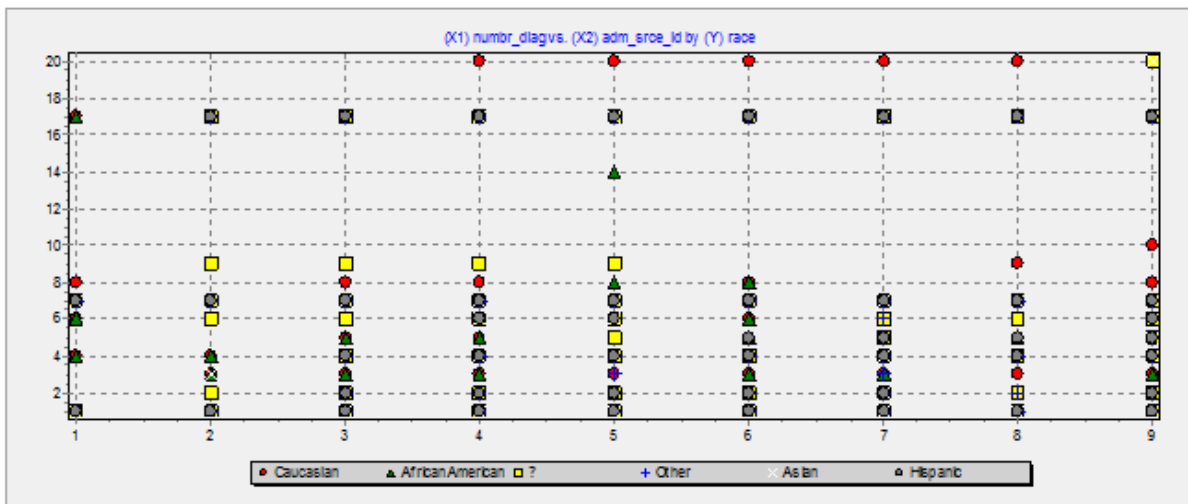


Figure 2: Scatter plot for least significant attribute diabetes diabetic medications with it supports attributes

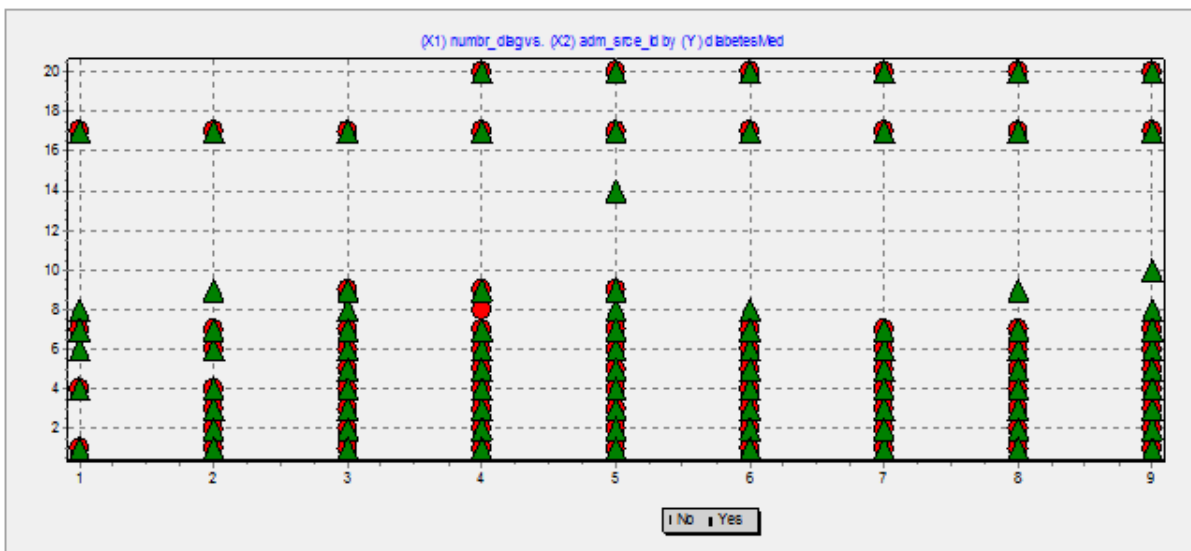
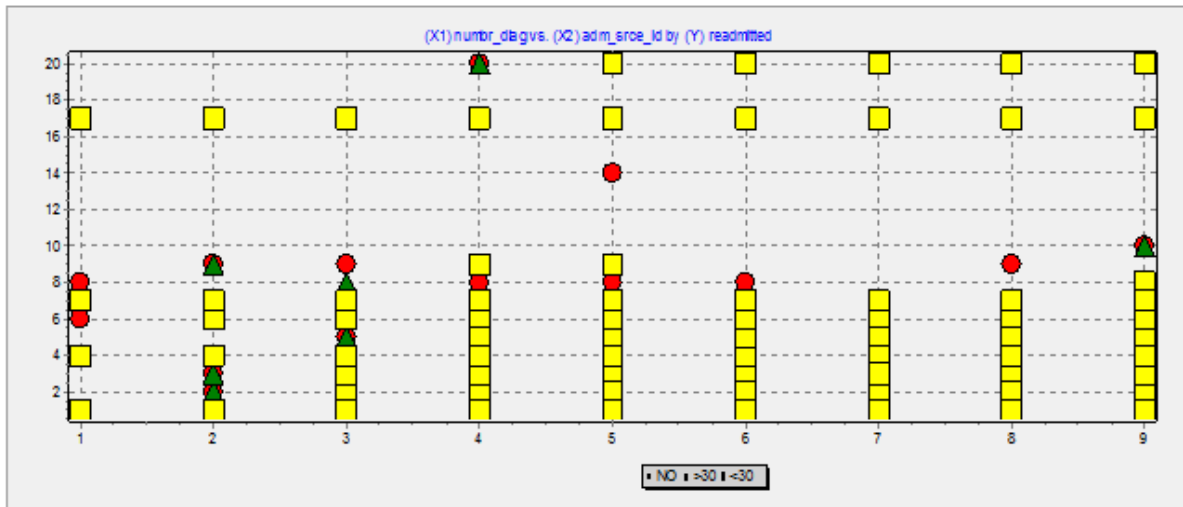


Figure 3: Scatter plot for least significant attribute readmitted with it supporting attributes



The use of PCA gives new set of axis of data with its variance. The variances are sorted as axes of a set. The highest variance of unit vector is considered as first axis and next highest variance is known as second axis and so on. The approximate strength of data in a dataset lies from stronger to weaker by sorting first axes to its last axes respectively. The table 3 consists of four axes with its correlation.

Table 3: Community estimates of attributes

Attribute	Axis_1		Axis_2		Axis_3		Axis_4	
	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)
-								
adm_type_id	0.0255	0 % (0 %)	0.6427	41 % (41 %)	0.4850	24 % (65 %)	0.1131	1 % (66 %)
dischrg_disp_id	0.3113	10 % (10 %)	-0.0743	1 % (10 %)	0.5553	31 % (41 %)	0.6277	39 % (80 %)
adm_srce_id	0.2591	7 % (7 %)	0.6450	42 % (48 %)	0.2483	6 % (54 %)	-0.4510	20 % (75 %)
tme_in_hsptl	0.5680	32 % (32 %)	-0.4192	18 % (50 %)	0.3190	10 % (60 %)	-0.0429	0 % (60 %)
num_emergcy	0.4427	20 % (20 %)	0.3794	14 % (34 %)	-0.5241	27 % (61 %)	0.1996	4 % (65 %)
num_inpatient	0.5772	33 % (33 %)	0.1496	2 % (36 %)	-0.4603	21 % (57 %)	0.2685	7 % (64 %)
numbr_diag	0.6367	41 % (41 %)	-0.2774	8 % (48 %)	0.1050	1 % (49 %)	-0.4718	22 % (72 %)
Var. Expl.	1.4218	20 % (20 %)	1.2535	18 % (38 %)	1.2045	17 % (55 %)	0.9465	14 % (69 %)

The eigen vector and factor scores states that the primary attributes in the data set is number of diagnosis, discharge disperse and admission source in table 4 respectively.

Based on the PCA results these attributes must be selected first and rest of attributes can be used as a supportive. The highest significant attributes are represented as primary components in first axes. Here locality of a dataset is considered as next highest significant attribute

Table 4: Eigen Vector and Factor scores of attributes

Attribute	Mean	Std-dev	Axis_1	Axis_2	Axis_3	Axis_4
Numbr_diag	7.091722	2.002187	0.533940	-0.247792	0.095712	-0.484919
dischrg_disp_id	4.080735	5.787179	0.261093	-0.066327	0.505947	0.645169
adm_srce_id	5.940078	4.511058	0.217327	0.576067	0.226213	-0.463568
tme_in_hsptl	4.521370	3.058125	0.476349	-0.374384	0.290695	-0.044127
adm_type_id	2.158984	1.554879	0.021404	0.574017	0.441880	0.116303
num_emergcy	0.145953	0.656894	0.371259	0.338880	-0.477525	0.205190
num_inpatient	0.603555	1.218315	0.484054	0.133597	-0.419397	0.275932

Figure 4: PCA axis Vs Axis 2 with race

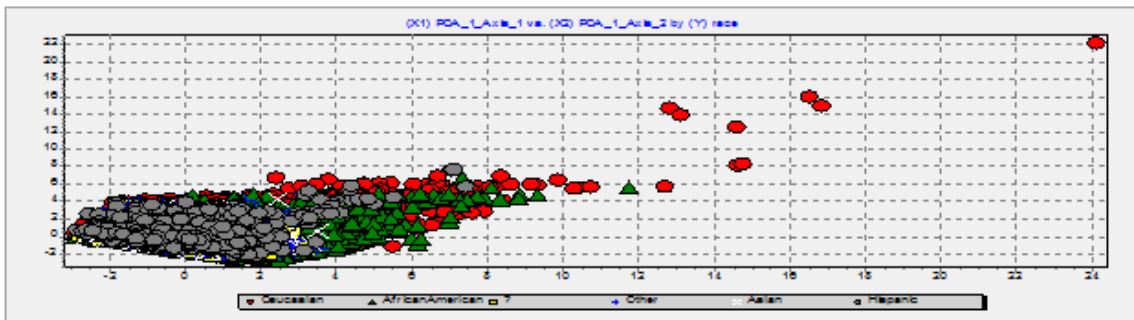


Figure 5: PCA axis Vs Axis 3 with race

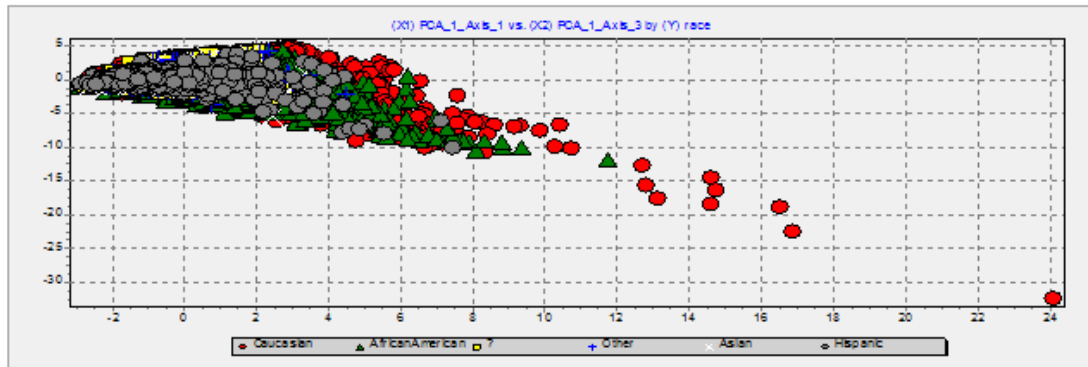
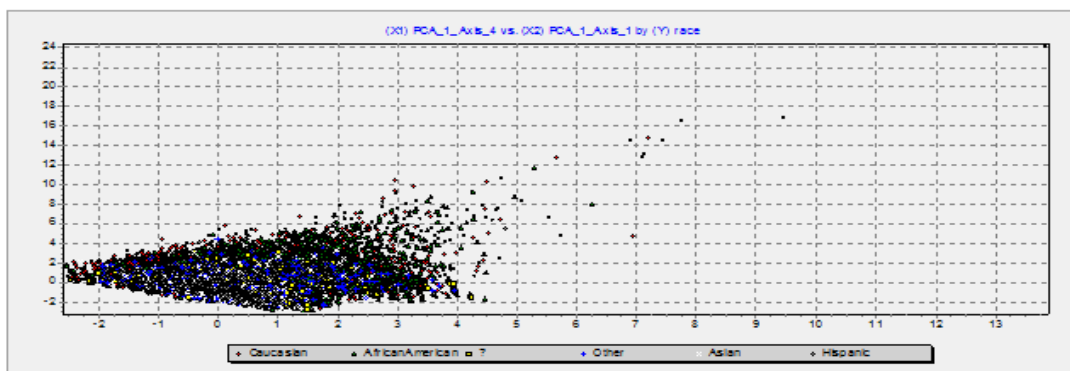


Figure 6: PCA axis Vs Axis 4 with race



The fourth axes are considered to be as least significant valued attributes. The scatter plots of a dataset shows variance in each axes. Note that the strength of fourth axes is very weak when compared with first axes and hence the first axes has highly dimensional categorical attribute of given data [1-13].

CONCLUSION

In this paper dimensionality reduction using PCA for very large data sets with multidimensional attributes has given better results. The results show that the big data analysis for clinical databases are categorised efficiently when significant attributes are selected earlier. The key point of this paper reveals that the extraction technique as well as computation methodology is always depends on the inherent nature of data. The final solution of this works proved that PCA technique with careful attribute selection gives most promising results for very large data bases.

REFERENCES

- [1] Pearson, K. 1901. "On lines and planes of closest fit to systems of points in space." *Philosophical Magazine* 2 (6): 559–72.
- [2] Hotelling, H. 1933. "Analysis of a Complex of Statistical Variables with Principal Components." *Journal of Educational Psychology* 24: 498–520.
- [3] Abdi, H., & Williams, L.J. (2010). "Principal component analysis." *Wiley Interdisciplinary Reviews: Computational Statistics*, 2: 433–459. doi:10.1002/wics.101.
- [4] Warmuth, M. K.; Kuzmin, D. (2008). "Randomized online PCA algorithms with regret bounds that are logarithmic in the dimension". *Journal of Machine Learning Research* 9: 2287–2320.
- [5] Golub, Gene H.; Van Loan, Charles F. (1996), *Matrix computations* (3rd ed.), Johns Hopkins University Press, Baltimore, Maryland, ISBN 978-0-8018-5414-9.
- [6] Arfken, G. "Eigenvectors, Eigenvalues." §4.7 in *Mathematical Methods for Physicists*, 3rd ed. Orlando, FL: Academic Press, pp. 229-237, 1985.
- [7] "Diabetes Blue Circle Symbol". International Diabetes Federation. 17 March 2006.
- [8] "About diabetes". World Health Organization. Retrieved 4 April 2014.
- [9] Dong Hyun Jeong, Caroline Ziemkiewicz, William Ribarsky and Remco Chang "Understanding Principal Component Analysis Using a Visual Analytics Tool"
- [10] E. Anderson and et al. *Principal Component Analysis*. the Society for Industrial and Applied Mathematics, third edition, 1999.
- [11] Asuncion and D. Newman. UCI machine learning repository, 2007.
- [12] Emily Mankin "Principal Components Analysis: A How-To Manual for R"
- [13] Jon Shlens, A TUTORIAL ON PRINCIPAL COMPONENT ANALYSIS Derivation, Discussion and Singular Value Decomposition.