International Journal of Statistics and Applications 2014, 4(6): 241-248
DOI: 10.5923/j.statistics.20140406.01

Identification of Outliers in Medical Diagnostic System Using Data Mining Techniques

V. Deneshkumar*, K. Senthamaraikannan, M. Manikandan

Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli, India

Abstract The outlier detection problem has important applications in the field of medical research. Clinical databases have accumulated large quantities of information about patients and their medical conditions. In this study, the data mining techniques are used to search for relationships in a large clinical database. Relationships and patterns within this data could provide new medical knowledge. The main objective of this paper is to detect the outliers and identify the influence factor in the diabetes symptoms of the patient using data mining techniques. Results are illustrated numerically and graphically.

Keywords Data mining, Outlier detection, Diabetes, PCA and refined method

1. Introduction

Outlier detection is a very important concept in the medical data analysis. The complex relationships that appear with regard to diabetic symptoms of the patient, diagnoses and behavior are the most promising areas of data mining. A data base may contain data objects that do not comply with the general behavior of the data. These data objects are outlier and the analysis of outlier data is referred to as outlier mining. Data mining is about finding new information from a large group of data. The problem of outlier detection for data mining is a rich area of research because the sequences are various types and outliers in sequences can be defined in multiple ways and hence there are different problem formulations. Most data mining methods discard outliers as noise or exceptions. The handling of outlier observations in a data set is one of the most important tasks in data pre-processing because of two reasons. First, outlier observations can have a considerable influence on the results of an analysis. Second, outliers are often measurement or recording errors, some of them can represent phenomena of interest, something significant from the viewpoint of the application domain. Some classical examples for inward procedures have given Hawkins [12] and Barnett and Lewis [2].

Factor Analysis is useful for understanding the underlying reasons for the correlations among a group of variables. The main application of factor analytic technique is to reduce the number of variables and to detect structure in the relationships among variables that classify variables. A small number of common factors are extracted so that these common factors are sufficient to study the relationship of original variables. The difficulties are having too many independent variables in such exercise that increases computational time to get solution, increased time in data collection, too much expenditure in data collection, presence of redundant independent variables, difficulty in making inferences, these can be avoided using factor analysis.

Gnanadesikan and Kettenring [6] discussed outlier detection in multivariate analysis and pointed out some of the problems. Hadi [7, 8] has addressed multivariate outlier detection problem to replace the mean vector by a vector of variable medians and to compute the covariance matrix for the subset of those observations with the smallest Mahalanobis distance. Penny and Jolliffe [18] have discussed a comparative study with six multivariate outlier detection methods. Bellazzi and Zupan [3] have discussed the new computational methods for data analysis and predictive modeling for clinical prediction problems. Hardin and Rocke [10] have developed a new method for identifying outliers in a one-cluster setting using an F distribution. They extended the method to the multiple cluster case which gives a robust clustering method in conjunction with an outlier identification method. They provided results of the F distribution method for multiple clusters which have different sizes and shapes.

Lavrac [15] has discussed the selected data mining techniques applied in medicine and in particular some machine learning techniques including the mechanisms that make them well suited for the analysis of medical databases. Wasan et al., [21] have discussed the impact of data mining techniques, including artificial neural networks for medical diagnostics. Ordonez et al., [16] have discussed the association rule used in the medical domain, where data sets

* Corresponding author:
vdenesh77@gmail.com (V. Deneshkumar)
Published online at http://journal.sapub.org/statistics

are generally high dimensional and small. Experiments focus on discovering association rules on a real data set to predict absence or existence of heart disease. Petoskey [19] has discussed outlier detection algorithms used in data mining systems and also explained their advantages and disadvantages. Parsons and Haque [17] have surveyed various subspace clustering algorithms along with a hierarchal algorithm by their characteristics. The proposed approach takes advantage of the unique characteristics of the data in this domain and provides a solution which is fast, scalable and produces high quality recommendations. Hodge [13] has surveyed the contemporary techniques for outlier detection and identified their respective motivations and distinguished their advantages and disadvantages. Ahmad and Dey [1] have proposed a modified k-mean algorithm for clustering mixed data sets. They also proposed a modified representation for the cluster center. The similarities of categorical attributes are calculated based on the proportion of items appearance. Koufakou and Georgiopoulos [14] have proposed a fast distributed outlier detection strategy intended for datasets containing mixed attributes. Chaira [4] has discussed a novel approach to intuitionistic fuzzy clustering using intuitionistic fuzzy set theory and showed its efficacy. This algorithm was tested on several CT scan brain images and the results were observed to be far better as compared to Type 2 fuzzy and conventional fuzzy C- means algorithm. Hauskrecht et al., [11] have developed and evaluated a data driven approach for detecting unusual patient management decisions using past patient cases stored in electronic health records (EHRs). This approach is capable of identifying clinically valid outliers at reasonably high true positive alert rates across a broad range of clinical conditions and patient management actions. Chrominski et al., [4] have discussed the use of different methods of outlier detection in medical diagnosis. Srimani et al., [17] have discussed outlier detection statistical methods in the application medical data set. The importance of outlier detection is due to the fact that outliers in data predict significant information in a wide variety of medical and other application domains. The information of medical data base has been chosen for this study, the information regarding the diabetic symptoms of the patient's details are collected. The objective is to detect the outliers and identify the influence factor of the diabetes patient using Factor analysis in which refined approach is used to detect the outliers in large clinical database.

2. Methodology

Factor analysis attempts to model the variance of an original set of variables by decomposing their variability into common and unique variance. The common variance or communality $\left(h_j^2\right)$ of a given manifest variable is the proportion of its variance that is accounted by the common factors' of the hypothesized latent structure. The unique variance $\left(1-h_j^2\right)$ of a given manifest variable is the proportion of variance that is unaccounted by the common factors [9]. The general factor model has the following form.

$$X_j = V_1F_1 + V_2F_2 + V_3F_3 + ... + V_mF_m + e_j \qquad (1)$$

where, X_j = The j[th] variable (j=1,2,3….p),

V_j = Factor loading of the j[th] common factor,

F_k = The k[th] common factor,

K = Number of common factors (k=1,2,3…m),

e_j = The unique variance of the j[th] variable

Principal Component Analysis

Principal component analysis is a multivariate statistical technique to identify or discover the underlying structure characterizing a set of highly correlated variables $\left(X_j\right)$. The principal components are extracted so that the first component accounts for the largest amount of total variation in the data, the second principal component accounts for the second largest amount of total variation and so on. A principal component which is a linear combination of set of variables has the following form

$$Z_j = W_1X_1 + W_2X_2 + W_3X_3 + ... + W_pX_p \qquad (2)$$

where,

Z_j = Linear composite of j[th] component (j =1, 2,…p number of variable).

W_j = The principal component loading of j[th] variable.

X_j = j[th] Variable.

In principal component analysis, the number of principal components extracted is the same as the number of variables describing the dimensionality of the data. If a data set is characterized by ten variables, principal component analysis of the data would generate ten principal components. The most important issue in principal component analysis is the number of principal components to retain for further analysis.

Varimax Rotation

A varimax rotation is a change of coordinates used in principal component analysis and factor analysis that maximizes the sum of the variances of the squared loadings.

$$\tilde{i}^* = \hat{l}_{ij}^* / \hat{h}_i \qquad (3)$$

Here $\hat{l}_{ij}^*$ is the loading of the i[th] variable on the j[th] factor after rotation, where $\hat{h}_i$ is the communality for variable i. The Varimax procedure, selects the rotation to find this maximum quantity. The sample variances of the standardized loadings for each factor, summed over the *m* factors.

$$V = \frac{1}{p}\sum_{j=1}^{m}\left\{\sum_{i=1}^{p}\left(\tilde{l}_{ij}^*\right)^4 - \frac{1}{p}\left(\sum_{i=1}^{p}\left(\tilde{l}_{ij}^*\right)^2\right)^2\right\} \qquad (4)$$

Computation Procedures for Refined Method

A refined procedure is applied, when both principal components and common factor extraction methods are used with factor analysis. Resulting factor scores are linear combinations of the observed variables shared between the item and the factor. The refined method is used to create factor scores, producing standardized scores similar to a Z-score method, where values range between -3.0 and +3.0.

Regression Scores

$$\hat{F}_{1xm} = Z_{1xn} B_{nxm} \quad (5)$$

where

n = Number of observed variables

m = Number of factors

F = The row vector of 'm' estimated factor scores

Z = The row vector of 'n' standardized observed variables

B = The matrix of regression of weights for the 'm' factors on the 'n' observed variables.

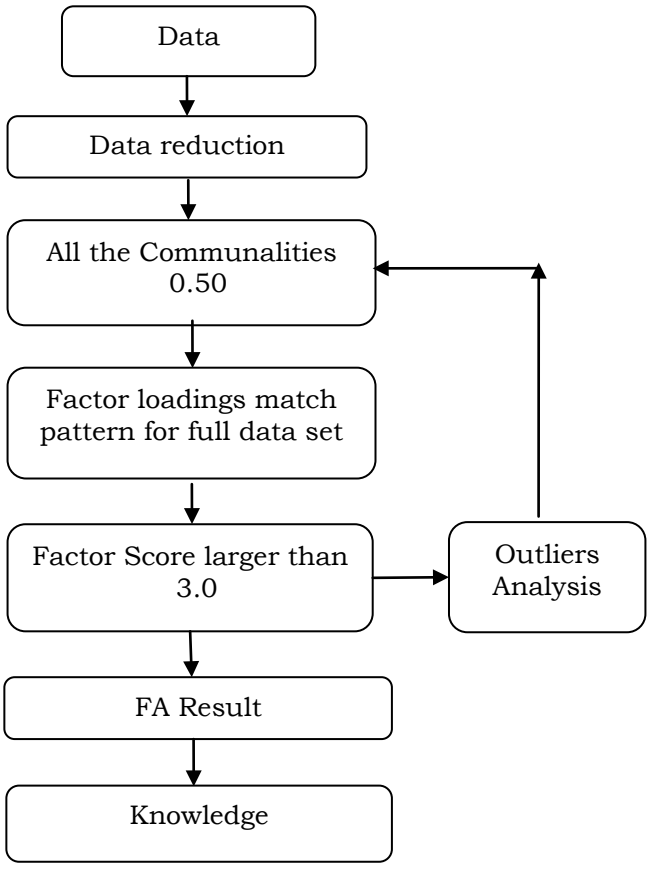


Figure 1. Flow Chart for the Model

Two Step Algorithm

In order to handle categorical and continuous variables, a Two-step cluster analysis procedure is employed. It uses likelihood distance measures that assume the variables in the cluster model are independent. Further, each continuous variable is assumed to have a normal distribution and each categorical variable is assumed to have a multinomial distribution. Empirical internal testing procedure is fairly robust to violations of both the assumption of independence and the distributional assumptions. The procedure of two-step cluster analysis algorithm can be summarized as follows

Step 1

The procedure begins with the construction of a Cluster Features (CF) Tree. The tree places the first case at the root in a leaf node, that contains variable information about that case. Each successive case is then added to an existing node or forms a new node, based upon its similarity to existing nodes and using the distance measure as the similarity criteria. A node that has multiple cases contains a summary of variable information about those cases. Thus, the CF tree provides a capsule summary of the medical data file.

Step 2

The leaf nodes of the CF tree are then grouped using an agglomerative clustering algorithm. The agglomerative clustering can be used to produce a range of solutions. To determine the best clusters, each of these cluster solutions is compared using Schwarz's Bayesian Criterion (BIC) or the Akaike Information Criterion (AIC) as the clustering criterion.

3. Result and Discussion

Outlier detection in the medical and public health domains typically work with patient records. The data can have outliers due to several reasons such as abnormal patient condition or instrumentation errors or recording errors. Thus the outlier detection is a very critical problem in this domain and requires high degree of accuracy. in this work we have analyzed the records which may have several different types of features such as patient age, height, weight and blood group.

The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy is an index used to examine the appropriateness of factor analysis. From the table 1 Kaiser-Meyer-Olkin measure of sampling adequacy value 0.632 indicates factor analysis is appropriate.

Table 1. KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.632
Bartlett's Test of Sphericity	Approx. Chi-Square	182.601
	df	28
	Sig.	.000

Bartlett's test of sphericity is a test statistic used to examine the hypothesis that the variables are uncorrelated in the population. In other words, the population correlation matrix is an identity matrix, each variable correlates perfectly with itself (r = 1) but has no correlation with the other variables (r = 0).

Table 2. Communalities

	Initial	Extraction
Polyphagia	1.000	.445
Polyuria	1.000	.533
Polydipsia	1.000	.654
Nocturia	1.000	.425
Weight loss	1.000	.664
Irritability	1.000	.381
Fatigue	1.000	.665
Blurred vision	1.000	.500

From table 2 communalities indicate the amount of variance in each variable that is accounted for initial and extraction values in diabetic symptoms for the patient. In initial column communalities are estimates of the variance in each variable accounted for by all components or factors in diabetic symptoms for the patient. In extraction column, communalities are estimates of the variance in each variable accounted by the components. It can be seen that table diabetic symptoms like Polydipsia, weight loss, and Fatigue are have high extraction values, which indicate that the extracted components represent the variables well.

Table 3. Total Variance Explained

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	1.942	24.275	24.275
2	1.314	16.420	40.696
3	1.011	12.642	53.337

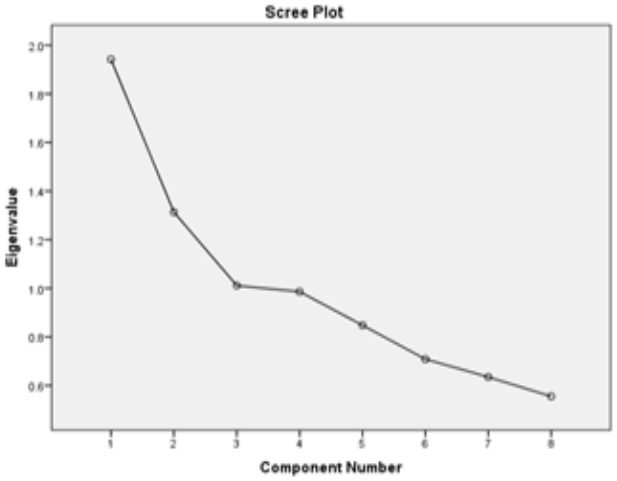


Figure 2. Scree Plot

From table 3 the first column shows the initial eigenvalues. The variance explained by the initial solution, extracted components in diabetic symptoms for the patient are displayed. The 2nd column gives the eigenvalue, or amount of variance in the original variables accounted by each component. The 3rd column gives the ratio, expressed as a percentage, of the variance accounted for by each component to the total variance in all of the variables.

The cumulative percentage for the second component is the sum of the percentage of variance for diabetic symptoms of the patient in the first and second components. They explain nearly 54% of the variability in the original ten variables considerably reduces the complexity of the data set by using these components, with 46% loss of information.

Figure 2 shows the scree plot that helps to determine the optimal number of components. The Eigenvalues of each component in the initial solution is plotted. The components 1, 2, and 3 on the steep slope are extracts and other components 4, 5,6,7,8 on the low slope contribute little to the solution.

Table 4. Rotated Component Matrix

	Component		
	1	2	3
Polyphagia	.653	.134	-.007
Polyuria	.697	.102	.190
Polydipsia	.797	-.126	.062
Nocturia	.027	-.121	.640
Weight loss	.079	.810	.037
Irritability	.226	.116	.563
Fatigue	-.171	.547	.580
Blurred vision	.427	.462	-.323

The table 4, shows the rotated component matrix of components. The first component is most highly correlated with Polydipsia. The second component is most highly correlated with weight loss. The third component is most highly correlated with Nocturia. So we focused on Polydipsia, weight loss and Nocturia for further exploration.

Table 5. Outlier Points

Patient ID	Fac1_1	Fac1_2	Fac1_3
71	1.52317	0.42783	4.83007
135	1.18491	-0.15401	-3.77214
254	1.64255	0.03768	-3.54834
78	1.64255	0.03768	-3.54834

From table 5 outliers are identified (because the computed factors scores a value is greater than ±3.0) in our analysis and then redo the analysis, omitting the cases that were outliers.

The model summary table indicates that five clusters were found based on the ten input features. The lower part of the figure 3 indicates the quality of the cluster solution. The silhouette measure of cohesion and separation is a measure of the clustering solution's overall goodness-of-fit.

It is essentially based on the average distances between the objects and can vary between -1 and +1.

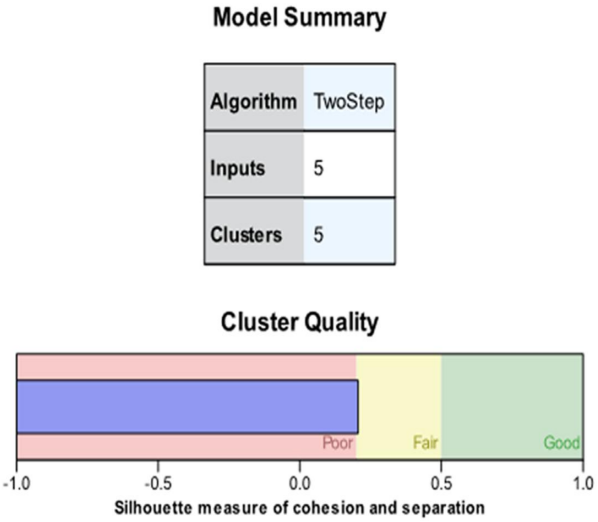
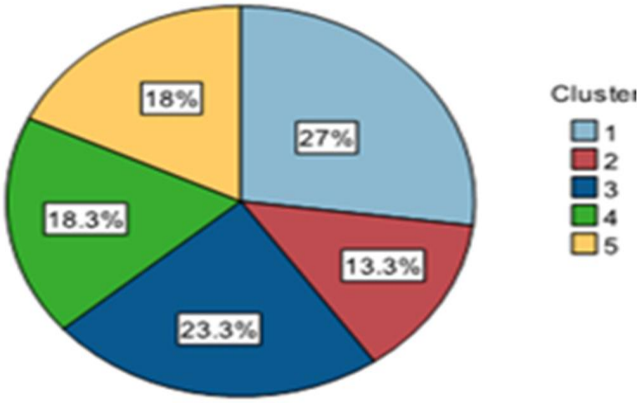


Figure 3. Model Summary and cluster quality

Specifically, a silhouette measure of less than 0.20 indicates a poor solution quality, a measure between 0.20 and 0.50 a fair solution, whereas values of more than 0.50 indicate a good solution In this case, the measure indicates a satisfactory solution.



Size of Smallest Cluster	40 (13.3%)
Size of Largest Cluster	81 (27%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	2.03

Figure 4. Model Summary

Cluster	1	3	4	5	2
Size	27.0%	23.3%	18.3%	18.0%	13.3%
Inputs	Polydipsia	Polydipsia	Polydipsia	Polydipsia	Polydipsia
	Weightloss	Weightloss	Weightloss	Weightloss	Weightloss
	Nocturia	Nocturia	Nocturia	Nocturia	Nocturia
	Age	Age	Age	Age	Age
	No.ofyrsaffected	No.ofyrsaffected	No.ofyrsaffected	No.ofyrsaffected	No.ofyrsaffected
Evaluation Fields	WeightKg	WeightKg	WeightKg	WeightKg	WeightKg
	Polyphagia	Polyphagia	Polyphagia	Polyphagia	Polyphagia
	Polyuria	Polyuria	Polyuria	Polyuria	Polyuria
	Fatigue	Fatigue	Fatigue	Fatigue	Fatigue

Figure 5. Cluster Distribution

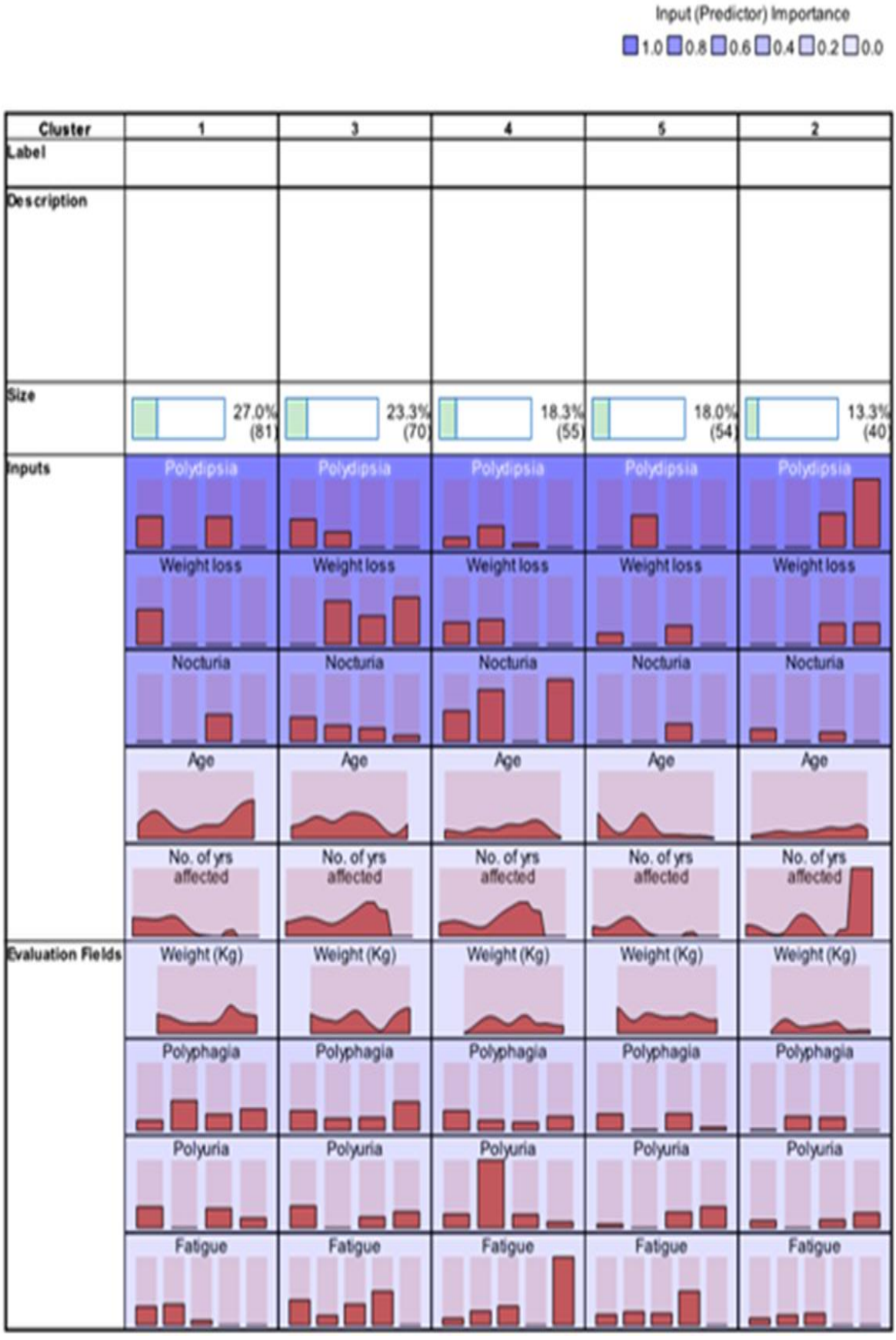


Figure 6. Cluster Structure

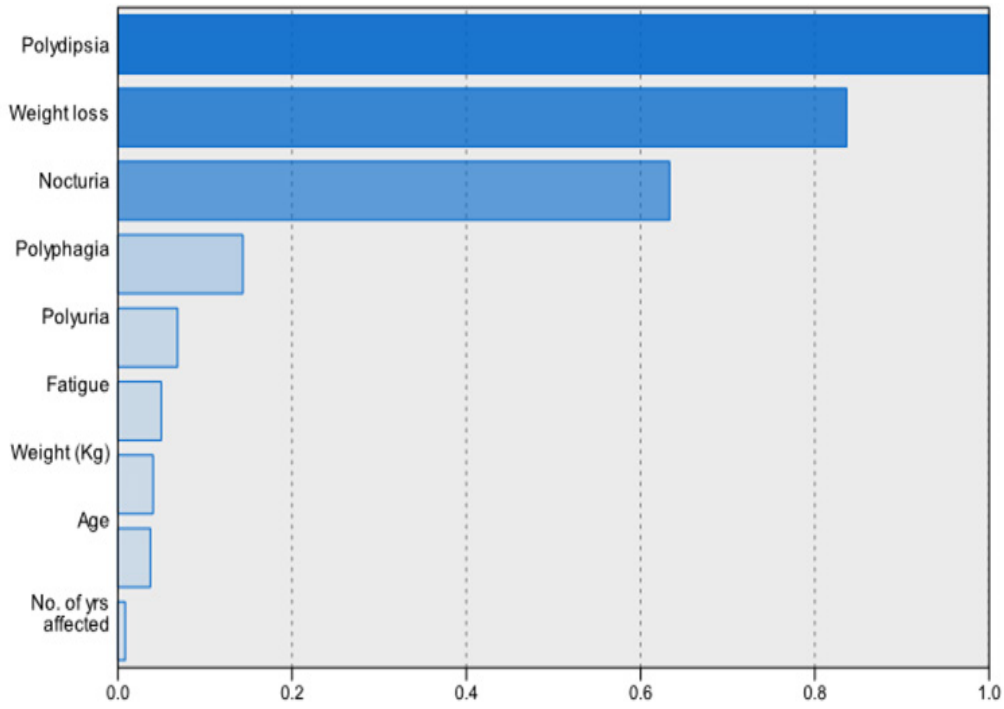


Figure 7. Cluster Preference

In figure 4 the cluster sizes view shows the frequency of each cluster. Hovering over a slice in the pie chart reveals the number of records assigned to the cluster. 27% of the records were assigned to the first cluster, 13.3% to the second, 23.3% to the third, 18.3% to the fourth, and 18% to the fifth.

Figure 5 shows the description of the five clusters, including their sizes. Furthermore, the output shows each clustering variables mean values across the five clusters as well as their relative importance. The cluster means suggest that the clusters are well separated. Darker shades denote the variables greater importance for the clustering solution with higher values. Comparing the results, we can see that polydipsia, weightloss and nocturia are the most important variable for each of the clusters, followed by age, no. of years affected, polyphagis and fatigue.

Figure 6 shows the different information of cluster, such as an overview of the cluster structure and visualization of the distribution of values for each field by cluster.

Figure 7 shows an over view of the variables' overall importance for the clustering solution, which provides the same result as the cluster-specific analysis. The model viewer provides us with additional options for visualizing the results or comparing clustering solutions.

4. Conclusions

The proposed outlier detection method for determining complex associations that influence medical outcomes by combining data mining with the patient record merits for further study. The complete sets of interdependent relationship are examined and the size of the medical data from ten variables to three components is reduced by using Factor analysis with principal components extraction. Three factors were identified for further exploration and this implies that detection of outliers in medical data applying data mining technique is more powerful and provided accurate extraction results. The process helps to the medical decision makers to provide better, consistent and efficient healthcare services.

ACKNOWLEDGEMENTS

The first author thanks the University Grants Commission, New Delhi for providing research fellowship under the scheme of BSRF to carry out this work.

REFERENCES

[1] Ahmad, L., and A. Dey (2007): K-Mean Clustering Algorithm for Mixed Numeric and Categorical Data. *Data & Knowledge Engineering*, Vol. 63, pp. 503-527.

[2] Barnett, V., and T. Lewis (1984): Outliers in Statistical Data, John Wiley & Sons, New York.

[3] Bellazzi, R., and B. Zupan (2008): Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*, Vol 77, No.2, pp. 81–97.

[4] Chaira, T. (2011): A Novel Intuitionistic Fuzzy C-Means Clustering Algorithm and Its Application to Medical Images. *Applied Soft Computing*, Vol.11, pp.1711–1717.

[5] Chrominski, K., and M. Tkacz (2010): Comparison of outlier detection methods in biomedical data. *Journal of medical Informatics & Technologies*, vol.16, ISSN 1642-6037.

[6] Gnanadesikan, R., and J. R. Kettenring (1972): Robust estimates, residuals, and outlier detection with multi-response data. *Biometrics*, Vol. 28, pp 81-124.

[7] Hadi, A.S., (1992): Identifying multiple outliers in multivariate data, *Journal of the Royal statistical Society*. Series B, Vol. 54, pp. 761-771.

[8] Hadi, A. S., (1994): A Modification of a Method for the Detection of Outliers in Multivariate Samples. *Journal of the Royal Statistical Society*. Series B, Vol.56, No.2.

[9] Hair, J. F., Black, J., Babin, W. C., Anderson, B. J., and R. L. Tatham (2006): Multivariate Data Analysis. 6th ed. New Jersey: Prentice Hall.

[10] Hardin.J and D.M. Rocke., 2002, Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator, Computational Statistics & Data Analysis, 44, 625 – 638.

[11] Hauskrecht, M., Batal, I., Valko, M., Visweswaran, S., Cooper, G.F., and G. Clermont (2013): Outlier Detection for Patient Monitoring and Alerting. *Journal of Biomedical Informatics*, Vol.46, pp.47–55.

[12] Hawkins, D.M., (1980): Identification of Outliers, Chapman and Hall.

[13] Hodge, V.J. (2004) A survey of outlier detection methodologies, Kluver Academic Publishers, Netherlands, January, 43.

[14] Koufakou, A., and M. Georgiopoulos (2010): A Fast outlier Detection Strategy for Distributed high-Dimensional Data sets with Mixed Attributes. Data Mining and Knowledge Discovery, Vol.20. pp. 259–289.

[15] Lavrac, N. (1999): Selected Techniques for Data Mining In Medicine, *Artificial Intelligence in Medicine*, Vol 16, No. 1, pp. 3–23.

[16] Ordonez, C., Ezquerra, N., and C. A. Santana (2006): Constraining and summarizing association rules in medical data, *Knowledge and Information Systems*, Vol.9 (3), pp.259–283.

[17] Parsons L., Haque, E. H. Liu, (2004): Subspace Clustering for high Dimensional Data: a Review, SIGKDD Explorations Vol.6(1), pp.90–105.

[18] Penny, K.I., and I.T. Jolliffe (2001): A Comparison of Multivariate Outlier Detection Methods for Clinical Laboratory Safety Data, *Journal of the Royal statistical Society*. Series D (The Statistician), Vol. 50, No. 3, pp. 295-308.

[19] Petrovskiy, M.I. (2003) Outlier Detection Algorithms in Data Mining Systems, *Programming and Computer Software*, Vol.29, No.4, pp.228–23.

[20] Srimani.P.K and Sanjay Koti.M., (2011): Application of Data Mining Techniques For outlier mining in Medical Databases, *International Journal of Current Research*, Vol. 33, No. 6, pp.402-407.

[21] Wasan, K.S., Bhatnagar, V., and H. Kaur (2006): The Impact of Data Mining Techniques on Medical Diagnostics, *Data Science Journal*, Vol 5, No.19. pp.119-126.