ISSN:0975-9646

Intelligent Data Analysis Using Data Mining Techniques

V.Ilango *1, Dr.R.Subramanian *2, Dr.V.Vasudevan *3

1Department of Computer Application, 2Computer Science, 3Information Technology

*Kalasalingam University, Krishnan oil, Srivilliputtur - 626 126, Tamil Nadu, India

Abstract-**The power of modern computing technology makes data gathering and storage easier. This leads to create new range of problems and challenges for data analysis. In this study a proposed approach based on clustering techniques for outlier detection is presented. At first EM-Cluster algorithm is performed to identify the missing values through which small clusters are formed. Then univariate outlier detection method is applied to identify outliers. The proposed approach gave effective results within optimum time and space when applied to synthetic data set.**

Keywords: **EM Cluster, Univariate outlier, Grubb test, Continuous variables**

I.INTRODUCTION

Data analysis is an area of high relevance in the fields of engineering and technology, where everyday prediction of rate is required, bioinformatics, medical science application, natural language processing, or customer relationship management. Clustering and outlier is now a significant research in computer science and other fields. The clustering algorithm was driven by biologist Sneath and Sokal in the 1963 in numerical taxonomy before being taken by the stastiscians [21]. In many clustering methods, clusters are often determined by estimating the location and dispersion of different sample groups within a given dataset [12]. In the literature outliers are found as a by-product of clustering algorithms that are neither a part of a cluster nor a part of the background noise; rather they are specifically points that behave very differently from the norm. [13][16][9][6] [19]. [8] proposed a new method that is a hyper clique-based data cleaner (HCleaner). These techniques are evaluated in terms of their impact on the subsequent data analysis, specifically, clustering and association analysis. According to [7] Outlier is defined as a data point that is very different from the rest of the data. Such a data point often contains useful information on abnormal behavior in the system that is characterized by the data. [10] Classified outlier detecting approach into two categories: classic outlier approach and spatial outlier approach. The classic outlier approach analyzes outlier based on transaction dataset, which can be grouped into statistical-based approach, distance-based approach, deviation-based approach and density-based approach. The spatial outlier approaches analyze outlier based on spatial dataset, which can be grouped into space-based approach and graph-based approach. [18] Discusses a cluster-outlier iterative detection algorithm, tending to detect the clusters and outliers in another perspective for noisy data sets. In this algorithm, clusters are detected and adjusted according to the intra-relationship within clusters and the inter-relationship between clusters and outliers, and vice versa.

The rest of the paper is organized as follows. Section 2 reviews related work in outlier detection. Section 3 describes the proposed method to detect missing values and outlier. Experimental results and their analysis are presented in Section 4 and finally, Section 5 concludes the paper.

II.RELATED WORK

In many engineering and application domains, noisy and missing samples often exist, causing negative affects on performance of data analysis techniques. The review of the related study has been discussed. [14] proposed PAM clustering algorithm to identify outliers. [1] has proposed context-sensitive clustering technique based on the Bayes decision theory to estimate an unsupervised way the statistical parameters of classes to be used in the Bayesian decision rule. The iterative procedure is based on the EM algorithm, which, starting from the estimates derived in the initialization steps, achieves the final values of the statistical parameters of classes to be used to accomplish the Bayesian classification. An investigation done by [2] reveals all data clustering algorithms have some ambiguity in some data when clustered. In web access navigation behavior EM-Clustering algorithm shows improved result when compared with k-mean algorithm [17]. The EM algorithm is introduced in order to estimate and improve the parameters of the mixture of densities recursively in color image segmentations [20]. The EM algorithm is presented in [9] to estimate the parameters corresponding to a probability density function when we have missing data. In this case the class labels are the missing data. [5] When items are missing the EM algorithm is a convenient way to estimate the covariance matrix at each iteration step of the BACON algorithm. A version of the EM algorithm for survey data following a multivariate normal model, the EEM algorithm (Estimated Expectation Maximization), is proposed. The combination of the two algorithms, the BACON-EEM algorithm, is applied to two datasets and compared with alternative methods. In [3] a search for outliers in two real data sets is shown. It is stressed that identifying outliers should not be done on the basis of asymptotical cutoffs derived under assumption of normality of the analyzed data. In [4] both dynamic programming approach (DPA) and grid-based pruning approach (GPA) are used for detecting outliers on uncertain data based on the definition of distance-based method

III DETECTION OF OUTLIERS USING CLUSTERING

There are a large number of clustering techniques to detect outliers. EM is chosen to cluster data for the following reasons among others .It has strong statistical basis and linear to data base size. EM is robust to noisy data and can accept the desired number of clusters as input. The EM algorithm proceeds by estimating the missing data (the E-Step) and then estimating the parameters of the model, via maximum likelihood (the M-Step) [4][5][6]. One way to identify univariate outliers is to convert all of the scores for a variable to standard scores. If the sample size is small (80 or fewer cases), a case is an outlier if its standard score is ±2.5 or beyond. If the sample size is larger than 80 cases, a case is an outlier if its standard score is ±3.0 or beyond [22] .This method applies to interval level variables, and to ordinal level variables that are treated as metric. Two sided Grubbs test is often used to evaluate measurements, coming from a normal distribution of size n, which are suspiciously far from the main body of the data. Grubbs' test is defined for the hypothesis: H0: There are no outliers in the data set. Ha: There is at least one outlier in the data set. The statistic is defined as:

$$G = \frac{|y_O - \bar{y}|}{s} \qquad (1)$$

With $\bar{y}$ and s denoting the sample mean and standard deviation, respectively, calculated with the suspected outlier included.

IV. EXPERIMENT AND RESULT

The data set used in this study was Hepatitis- medical data set obtained from uci repository. It has 20 attributes and 155 instances consisting of 6 continuous attributes and 14 discrete attributes. EM algorithm was implemented and four clusters were constructed with maximum likelihood value (-26.58564). Data was analyzed and 13 missing values were identified and hence they were eliminated as shown in Table 2. Finally 142 samples were considered for analysis. Four clusters were formed as shown in Figure.1. Table.3 shows the cluster description, the mean and standard deviation for the corresponding clusters. With this univariate outlier detection, Grubb test was experimented on continuous variables to find the outliers as shown in Figure.2. Table.4. Shows total number of detected outliers, respective observations and their values. Optimum time and space taken for computation is shown in Table.1.

TABLE.1. COMPUTATIONAL DETAIL

Label	Count	Original Value	Final	Missing Values
No of attribute	20	…	…	…
Continues attributes	6	…	…	…
Discrete attributes	14	…	…	…
Computational time	17ms	…	…	…
Allocated memory	30kb	…	…	…
No of Instances	…	155	142	13

TABLE.2. MISSING VALUES DETAILS

Missing Instances	Missing Attributes
4	Steroid
31,87,100, 111,118, 132	Liver big, Liver Firm
40	Liver big, Liver Firm, Spleen Palpable,Spider,Ascities,Varcies
54	Fatigue, Malaise, Anorexia, Liver big, Liver Firm, Spleen Palpable,Spider,Ascities,Varcies
69	Liver big, Liver Firm, Spleen Palpable,Spider,Ascities,Varcies
79	Spleen Palpable,Spider,Ascities,Varcies
137	Liver big, Liver Firm, Spleen Palpable,Spider,Ascities,Varcies
138	Liver Firm

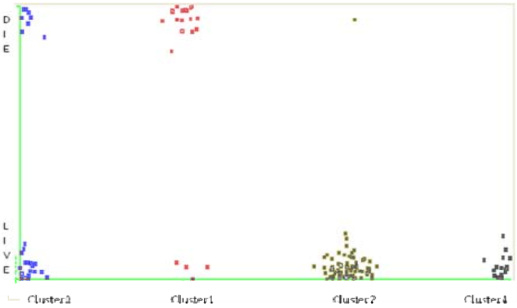

Figure.1. Cluster wise Samples

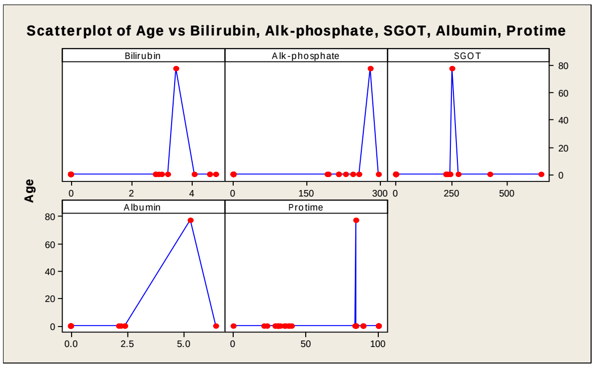

Figure 2. Outlier Detail Variable wise

TABLE.3. DESCRIPTION OF CLUSTER

Att - Desc	Test value	Group Mean (StdDev)
Cluster=cluster0 Examples 28(19.7%)		
ALK_PHOSPHATE	1.03	154.57 (62.31)
AGE	0.46	46.54 (14.79)
BILIRUBIN	0.31	1.77 (1.00)
SGOT	0.26	104.92 (76.78)
PROTIME	-0.21	58.36 (11.59)
ALBUMIN	-0.31	3.63 (0.47)
Cluster=cluster1 Examples 22(15.5%)		
BILIRUBIN	1.16	2.79 (2.15)
ALK_PHOSPHATE	0.14	111.98 (39.81)
AGE	0.12	42.45 (9.21)
SGOT	-0.05	79.18 (51.20)
PROTIME	-0.92	45.73 (15.50)
ALBUMIN	-1.38	2.97 (0.48)
Cluster=cluster2 Example 76(53.5%)		
ALBUMIN	0.44	4.10 (0.49)
PROTIME	0.27	66.90 (16.98)
AGE	-0.11	39.70 (11.76)
SGOT	-0.31	58.02 (42.55)
BILIRUBIN	-0.42	0.91 (0.26)
ALK_PHOSPHATE	-0.42	85.52 (24.53)
Cluster=cluster3 Example 16(11.3%)		
SGOT	1.07	171.13 (167.13)
ALBUMIN	0.33	4.03 (0.36)
PROTIME	0.33	67.84 (18.70)
ALK_PHOSPHATE	0.01	105.94 (48.92)
BILIRUBIN	-0.17	1.21 (0.59)
AGE	-0.48	35.19 (7.65)

TABLE.4 OUTLIER DETAILS VARIABLE WISE

Variable	#Observation	Outlier	Values
Age	3	1	78
Bilirubin	59,64,68,86,91,94,97,108,111,112,118,120,122,136,138	15	3.5,4.1,2.8,4.6,3,4.8,4.6,3.2,2.9,2.8,4.6,8,4.2,4.2,7.6
Alk-phosphate	30,35,45,59,85,97,103,125,131	9	280,194,191,215,230,215, ,256,243.295
SGOT	11,45,72,77,96,100,101,115,123,138	10	249,420,242,224,227,648,225,231,278,242
Albumin	70,98,99,118,122,134	6	5.3,2.1,6.4,2.4,2.2,2.4
Protime	10,18,21,22,25,27,28,30,35,37,39,41,46,51,57,58,59,77,78,84,88,90,94, 102,104,113, 117,118,120,121,125,129,139,133	34	85,85,39,100,100,36,100,40,90,100,21,100,85,100,100, 90,29,100,100,84,38,100,31,31,29,23,100,32,30,0,90,31,85,35

V.CONCLUSION

In this study a proposed approach based on clustering techniques for outlier detection is presented. EM clustering algorithm is performed and missing values are identified. Small clusters are formed. Then univariate outlier detection method is applied and reasonable amount of outliers are identified. But outliers are not removed as it may produce sensible information for decision making. Based on continuous variable the experimentation was carried out. The proposed approach gave effective results when applied to synthetic data set within optimum time and space. For the future we would perform using categorical variable of real data with supervised machine learning techniques.

REFERENCES

[1] Aggarwal CC et al (1999) Fast algorithms for projected clustering. In: Proceedings of ACM SIGMOD, pp 61–72

[2] Agga rwal CC, Yu P (2000) Finding generalized projected clusters in high dimensional spaces. In: Proceedings of ACM SIGMOD, pp 70–81

[3] Anna Bartkowiak, "Outliers in Biometrical Data – Two Real Examples of Analysis", 2009,International Conference on Biometrics and Kansei Engineering, P: 1-6

[4] Bin Wang, Gang Xiao, Hao Yu, Xiaochun Yang, "Distance-Based Outlier Detection on Uncertain Data", IEEE Ninth International Conference on Computer and Information Technology- 2009,P:293- 298

[5] Cédric Béguin And Beat Hulliger, "The BACON-EEM Algorithm For Multivariate Outlier Detection In Incomplete Survey Data", Survey Methodology, June 2008 ,Vol. 34, No. 1, Pp. 91- 103statistics Canada, Catalogue No. 12-001-X

[6] Guha S, Rastogi R, Shim K (1998) CURE: an efficient clustering algorithm for large databases.In: Proceedings of ACM SIGMOD, pp 73–84

[7] Hawkins, D. (1980). Identification of Outliers. Chapman and Hall. London.

[8] Hui Xiong, Gaurav Pandey, Michael Steinbach, Vipin Kumar, " Enhancing Data Analysis with Noise Removal" , IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 3, MARCH 2006 ,P:304-319

[9] Iuliana F Iatan, "The Expectation- Maximization Algorithm: Gaussian Case", IEEE International Conference on Networking and Information Technology", 2010, 978-1-4244-7578-0,Pp:590-593

[10] Jingke Xi, "Outlier Detection Algorithms in Data Mining", Proceeding of Second International Symposium on Intelligent, P.94-97, 2008

[11] Kaufman, L. and Rousseeuw, P.J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

[12] Lotfi A. Zadeh. Fuzzy logic, neural networks and soft computing, November 1992. One-page course announcement of CS 294-4, Spring 1993, University of California, Berkeley.

[13] Mayank Tyagi et.al., "A Context-Sensitive Clustering Technique Based on Graph-Cut Initialization and Expectation-Maximization Algorithm", IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, VOL. 5, NO. 1, JANUARY 2008, pp: 21-25

[14] Moh'd Belal Al- Zoubi, "An Effective Clustering-Based Approach For Outlier Detection, European Journal Of Scientific Research Issn 1450-216x Vol.28 No.2 (2009), Pp.310-316

[15]Norwati Mustapha. et.al., "Expectation Maximization Clustering Algorithm for User Modeling in Web Usage Mining Systems" , European Journal of Scientific Research, ISSN 1450-216X, Vol.32 No.4 (2009), pp.467-476

[16]Osama Abu Abbas, "Comparison between data clustering algorithms", The International Arab Journal of Information Technology, Vol.5, No.3, July 2008

[17] Rama.B et.al, "A Survey on clustering-Current status and challenging issues" , International Journal on Computer Science and Engineering , Vol. 02, No. 09, 2010, 2976-2980

[18] Yong Shi, "Detecting Clusters and Outliers for Multi-Dimensional Data", 2008 International Conference on Multimedia and Ubiquitous Engineering, P:429-432

[19] Zhang T, Ramakrishnan R, LivnyM(1996) BIRCH: an efficient data clustering method for very large databases. In: Proceedings of ACM SIGMOD, pp 103–114

[20] Zhi-Kai Huang and De-Hui Liu, "Segmentation of Color Image Using EM algorithm in HSV Color Space", Proceedings of the 2007 International Conference on Information Acquisition, July 9-11, 2007, Jeju City, Korea, pp:316-319.

[21] http://en.wikipedia.org/wiki/Numerical_taxonomy

[22] Hair et al, "Multivariate Data Analysis", Pearson education, fifth edition, 2003.