International Scientific Journal of Contemporary Research in Engineering Science and Management.
ISSN: 2456-1134
Vol.No:03
doi:

Received:**15/12/2018** Revised:**15/01/2019** Accepted: **15/02/2019**

Intelligent Data Analysis in Healthcare industry using Novel Decision Support System

Dr.S.Mohan kumar,

New Horizon College of Engineering, Research Centre,Bangalore,India.
E-Mail:drsmohankumar@gmail.com

Abstract - Extensive amounts of knowledge and data stored in business databases require the development of specialized tools for storing and accessing of data, data analysis, and effective use of stored knowledge and data. This paper focuses on methods and tools for intelligent data analysis, aimed at narrowing the increasing gap between data gathering and data comprehension. The paper sketches the history of research that led to the development of current intelligent data analysis techniques, discusses the need for intelligent data analysis in business, and proposes a decision support system of intelligent data analysis methods. The scope of the paper covers temporal data abstraction methods and data mining methods. 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**

1. Introduction

Decision support is a crucial function for decision makers in many industries. Typically, decision support systems help decision-makers to gather and interpret information and build a foundation for decision-making. Such systems may range from simple software systems to complex knowledge-based and artificial intelligence systems. Decision support systems can be database-oriented, spreadsheet-oriented or text-oriented in nature. In healthcare, clinical decision support systems (CDSS) can play a significant role. Clinical decisions that are routinely taken by healthcare service providers are often based on clinical guidance and evidence-based rules derived from medical science. However, intelligent decision support systems (IDSS), through the interpretive analysis of large-scale patient data with intelligent and knowledge-based methods, “allow doctors and nurses to quickly gather information and process it in various ways in order to assist with making diagnosis and treatment decision” [1]. IDSS can be applied in healthcare in diverse areas such as the examination of real-time data from diverse monitoring devices, analyses of patient and family history for the purpose of diagnosis, reviews of common characteristics and trends in medical record databases and many more areas.

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 [6]. 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 ery 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. [18] 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.

INTELLIGENT DECISION SUPPORT

An IDSS induces specific domain knowledge from raw data by identifying and extracting strategically useful information patterns from this data, thus making the extracted patterns understandable and usable for decision-making. IDSS, unlike DSS, “allows for supporting a wider range of decisions including those with uncertainty” [1]. IDSS, in addition to giving recommendations, may also contribute estimates of the level of confidence in the recommendations it gives.

IDSS can handle complex problems, applying domain-specific expertise to assess the consequences of executing its recommendations. Decisions supported by IDSS also tend to be more consistent, timely and better managed in terms of managing uncertainty in the outcomes. The justification of outcomes provided by an IDSS is particularly significant if it allows clinical experts to validate the explanations provided by the IDSS [1].

Managing knowledge in healthcare organizations to aid clinical decision-making requires transforming information into actionable intelligence that can be interpreted by different functional workgroups within the organization. This is demonstrated in Figure 1, a representation of the healthcare knowledge cycle from Patel et al [4], which shows how artificial intelligence can be used to analyze healthcare data and generate a representation of knowledge that can in turn be used for information and process modeling.

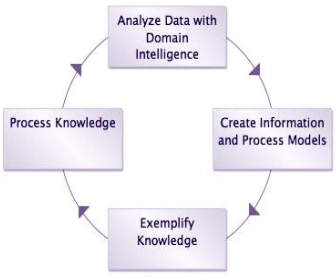


Figure 3: The knowledge cycle implemented with AI methods and tools

ways in clinical decision-making at both the individual patient

level and the population level. For example:

Diagnose by regularly interpreting and monitoring patient data. An IDSS can implement rules and patterns for individual patients, based on clinical parameters, and raise warning flags when such rules are violated. These flags can lead to clinical interventions that save lives.

- Help chronic disease management through establishing benchmarks and alerts. For chronically ill patients, a deviation noticed by an IDSS in, say, a blood test reading from a diabetic patient could result in an intervention before the patient gets into difficulty.
- Help public health surveillance by detecting pandemic diseases or in surveillance of chronic diseases. In case of a pandemic, an IDSS can interpret data and predict possible future spread of the disease.
- Additionally, IDSS can perform regular clinical decision support functions like preventing drug-drug interactions.Even if not noticed by the prescribing physician, an IDSS can spot incompatibilities between prescribed medications and/or dosages for the patient.

Architecture of an IDSS Model

The IDSS model described here combines an association rule-generator algorithm based on data mining of a knowledge base (KB) with an artificial neural network (ANN) system [2]. The system is capable of building domain knowledge from existing datasets and applying this knowledge to solve clinical problems. As data mining extracts domain-specific knowledge from organizational databases, it also enhances the process of knowledge acquisition. Neural networks can learn patterns from large volumes of data and use the knowledge thus extracted to help solve problems [2].

The solution discussed here has been adapted from a similar architecture presented by Viademonte et al. [2]. In the context of the primary healthcare system, the system creates and tracks patient profiles and then uses the patterns it recognizes from the data to identify unusual test readings and trigger alerts for possible intervention. It can also assist in diagnosing certain diseases based on a set of observed symptoms and suggest recommendations. The IDSS retrieves information learned in the past, creates domain knowledge from recalled information and translates it into “new” domain knowledge, to serve as a predictive tool. It is basically a hybrid system for applying descriptive and predictive models for intelligent decision-making [2][3].

This system allows raw data to be retrieved from data bases and processed into data models. These data models support descriptive methods that are stored in knowledge bases. Subsequently, predictive methods (based on neural network models) are generated that produce predictions [2].

The IDSS architecture under discussion can operate either through data mining for knowledge acquisition or through a neural network-based system operating as an advisory system. While the data mining technology offers expertise, the ANN-based system acquires knowledge through learning and reasoning as well at the intuitive user interface level [2]. At the data level, the system depends on a master data warehouse that combines relevant data repositories, case bases and knowledge bases. The elements of the architecture include [2]:

- a decision-oriented data repository, such as a data warehouse
- case bases
- inductive algorithms for data mining (descriptive method)
- knowledge bases
- an intelligent advisory system (predictive method)

Through a training process, data mining algorithms can be applied to multiple data bases to build multiple descriptive models that are stored in knowledge bases. As neural networks are applied to these descriptive models, predictive models are generated. Domain specific cases or case bases and their corresponding data models are created by extracting data from the data warehouse. The process ensures data consistency within that domain [3]. A “case” in this discussion represents an instance of a problem within the specific domain and falls into a specific well-defined class consisting of attributes and values [2][3].

Once data mining has been successfully employed to extract relevant relationships from the case bases, association rules are applied to produce general knowledge that is stored in the knowledge base. In the medical field, specific clinical cases or practice guidelines can be used as case bases from which data can be mined to produce clinical knowledge for generating descriptive clinical features or for decision support functions [2][3].

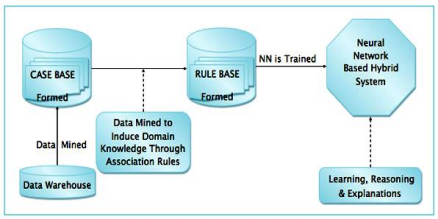


Figure 4. Main building blocks of the IDSS Model

The dashed lines in Figure 4 symbolize processes and the solid lines with arrowheads symbolize data flows between components. The data warehouse consists of pre-processed historical data that are mined. Subsequently, cases are selected, extracted and stored in case bases [2].

2.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

3.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:

$$G = \frac{|y_O - \bar{y}|}{s}$$

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

4. 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

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. The IDSS model described in this article is capable of learning, generalizing and self-organizing in order to recognize complex patterns and assist in decision support. The case study of Jane, the 41-year-old diabetic patient, indicates that when self-monitoring data and test data at each patient visit are available to a physician using an EMR with IDSS support, the physician would be able to make better decisions. In the future, uploading self-management monitoring data automatically from patient personal health records to family physician electronic medical records may become the norm for chronically ill patients, and the intelligent decision support system discussed here would be able to play an important role in providing improved patient care.

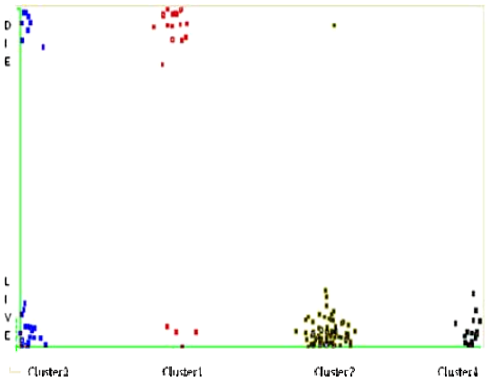
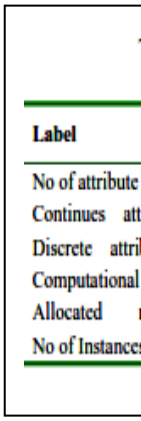


Figure 2 Cluster wise Samples



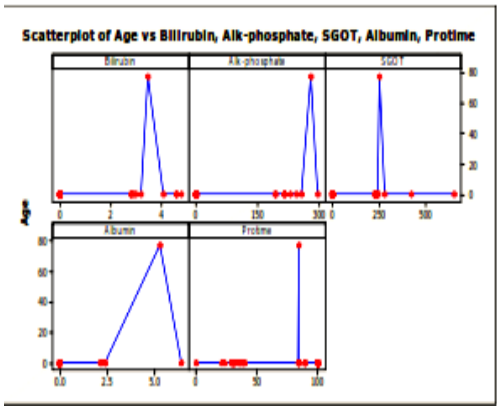


Figure 1 Outlier Detail Variable wise

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

TABLE.3. DESCRIPTION OF CLUSTER

Cluster=cluster0 Examples 28(19.7%)		
Att - Desc	***Test value***	***Group Mean (StdDev)***
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%)		
Att - Desc	***Test value***	***Group Mean (StdDev)***
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%)		
Att - Desc	***Test value***	***Group Mean (StdDev)***
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%)		
Att - Desc	***Test value***	***Group Mean (StdDev)***
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)

5.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.Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias De Amorim, and Serge Fdida. From popularity prediction to ranking online news. Social Network Analysis and Mining, 4(1):1–12, 2014

4. Anna Bartkowiak, “Outliers in Biometrical Data–Two Real Examples of Analysis”, 2009,International Conference on Biometrics and Kansei Engineering, P: 1-6

5.Basu, R., Fevrier-Thomas, U., Sartipi, K., “Incorporating hybrid CDSS in primary care practice management,” McMaster eBusiness Research Centre, November 2011.

6.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.

7.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

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

9.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.

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.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

13.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.

14.Viademonte and F. Burstein, “From knowledge discovery to computational intelligence: A framework for intelligent decision support systems,” Chapter 4 in “Intelligent Decision-making Support Sytems,” Springer-Verlag London Limited, pp. 57-78. 2006

15.V. L. Patel, E. H. Shortliffe, M. Stefanelli, P. Szolovits, M. R. Berthold, R. Bellazzi and A. Abu-Hanna, “The coming of age of artificial intelligence in medicine,” Artificial Intelligence in Medicine, Vol. 46, No. 1, pp. 5-17, 2009.

16.V.Ilango,“A Five Step Procedure for Outlier Analysis in Data Mining-Survey”, European Journal of Scientific Research, ISSN 1450-216X Vol.75 No.(2012), p. 327-339.

17.V.Ilango, “Intelligent Data Analysis Using Data Mining Techniques”, International Journal of Computer Science and Information Technologies, Vol.2 (4) ,2011,p.1420-1422.

18. V.Ilango, “Statistical Data Mining Approach with Asymmetric Conditionally Volatility Model in Financial Time Series Data”, International Journal of Soft Computing, ISSN: 1816-9503(Print), Vol.8 (4), December 2013.

19.V.Ilango, “Outlier detection and influential point observation in linear regression using clustering techniques in financial time series data”, Journal of Theoretical and Applied Information Technology, E-ISSN 1817-3195 Vol. 49. No. 2 – 2013,p.536-549.

20.V.Ilango,“Outliers Detection for Regression using K-Means and Expected Maximization Methods in Time Series Data”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 9,p.1052-1060, September 2013.