International Journal of Electrical, Electronics and Computer Engineering **6**(1): 150-157(2017)

ISSN No. (Online): 2277-2626

An Hybrid Data Mining Approach to detection and classification of Health Care Data

Madiha Akhtar* and Rajvardhan Singh Parihar*

**Department of Computer Science & Engineering,*
IASSCOM Fortune Institute of Technology Bhopal, (Madhya Pradesh), INDIA

(Corresponding author: Madiha Akhtar)
(Received 30 March, 2017 Accepted 25 May, 2017)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Due to the technological development in health care informatics, digitalizing health records and telemedicine has resulted in rapid growth in the production of enormous amount of health data which clutches multifaceted information relating to patients and their medical circumstances. To extract and discover the useful information from outsized dataset data mining techniques uses which is applied on medicinal data to enhance the services of the health care division. Outlier detection is one of the data mining techniques which can be used in various field of medical division to analyze outsized dataset but the conventional outlier detection is not much efficient for large scale and high dimensional health care data. This paper proposes a hybrid outlier detection method namely ID3 with KNN and GA to develop effective outlier detection. The proposed ID3-KNN & GA agrees to the case categorization quality character (CCQC) with the medical quality evaluation model and utilizes the attribute overlapping rate (AOR) algorithm for data classification and dimensionality reduction. To calculate the performance of the pruning operations in ID3-KNN & GA, we perform widespread experiments on accuracy and specificity parameters. The experiment consequences show that the ID3-KNN&GA method outperforms the k-nearest neighbor (KNN) in terms of the accuracy and efficiency.

Keywords: K-Nearest Neighbor; ID3; GA; health care; outlier detection; attribute overlapping rate; case categorization quality character

I. INTRODUCTION

Due to the extensively growth in technological industry, it makes our task easier to gather health care data from various sources such as computerized patients record, blood pressure monitors, electronics scales and wearable etc. [1-2]. Health care data is extremely challenging task not only because of its volume but also the diversity of types of information and high dimensions. For each and every record, the dimension can be vary from hundreds to thousands, which entails the hospital patient, doctor, medicine information, medicinal equipment etc. As patients will utilize various types of tablet or drugs, the dimensionality will enhance as the increase of the number of records. Such outsized data in health care (e.g., EMR - Electronic Medical Record) is so individually complex that traditional approaches to analyze and reveal the outliers will not work efficiently. Outlier is an example which is not practically equivalent to as for whatever remains of the examples in the dataset. Contingent upon the application domain, outliers are of abnormal intrigue. Exposing outliers may prompt to the disclosure of genuinely surprising conduct and help stay away from wrong conclusions and so forth which gives us separated and cleared information to work on. Occasionally, outliers are just noise or "polluted data", but more habitually, they corresponds exceptional and meaningful information. For instance, a breast cancer detection system might consider inliers to represent healthy patient and outlier as a patient with possibility of breast cancer. Fig. 1 illustrates outliers in a simple 2-dimensional data set. The data has two normal regions, N1 and N2, since most observations lie in these two regions. Points that are sufficiently far away from the regions, e.g., points o1 and o2, and points in region O3, are outliers.

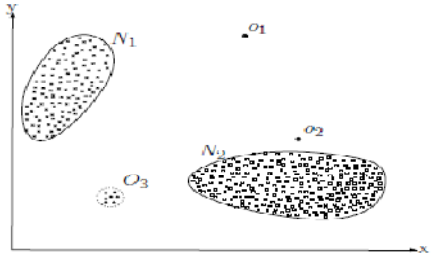

Fig. 1. A simple example of outliers in a 2-dimensional data set.

The types of outliers can be classified into 3 various classes namely:

- Point outliers which deals with multidimensional data types
- Contextual outliers based on the dependency-oriented data types such as discrete sequences, time-series, data and graphs. Every instance to a context is defined using the attributes such as contextual attributes and behavioral attributes.
- Collective outliers states that the individual data instance is not an outlier whereas a collection of related data may form an outlier

The massive number of unsupervised, semi supervised and supervised algorithms are found in the literature for outlier exposure. These algorithms again can be classified to classification-based, clustering-based, nearest neighbor based, density based, information theory based, spectral decomposition based, visualization based, depth based and signal processing based techniques.

Outlier detection has attracted more research in data mining, health care fraud, intrusion detection analysis and others. The common outlier detection algorithms include distance-based, statistical-based, density based, depth-based and cluster-based detection approaches. In distance-based outlier detection, K-Nearest Neighbor (KNN) is a simple and efficient outlier detection method. The outlier is defined as the data whose distance to its k^{th} nearest neighbor is larger than the outlier threshold. The limitation of such a method is its efficiency when the data is not uniformly distributed and its size is large. In density-based outlier detection, Local Outlier Factor (LOF) method is a local detection method. The outlier is defined as the data whose local density is significantly smaller than its neighbors. As a local detection approach, LOF can deal with non-uniform distribution of the data, but may not be well suitable for high data dimension and large data sparsity. However, for health care data, generally, the distribution is non-uniform; the dimension is high; the density is small and the size is large. The traditional outlier detection algorithms cannot effectively handle such health care outsized data with high dimension and large-size as well as sparse and unequal distribution.

In this paper, we propose a novel hybrid outlier detection method ID3, K-Nearest Neighbor & GA methods to effectively reduce and prune data dimensionality for outlier detection. With ID3-KNN & GA, Case Classification Quality Character (CCQC) [11] is adopted as the evaluation model. The three main evaluation indexes, medical model, medical defect and medical trend are the basis for dimension reduction. Attribute overlapping rate (AOR) algorithm is then used for dimensionality reduction and data classification. AOR reduces the data dimension without losing important information, and divides the data into small data sets so that it can be performed on the MATLAB platform in parallel. To improve the detection accuracy and efficiency, a series of pruning operations are then conducted.

The rest of this paper is organized as follows. Section II describes the types of outlier and its detection with related techniques. Section III reviews the related work in outlier detection and medical quality evaluation. The ID3-KNN&GA method is proposed in Section IV. The experimental results with real-world data are presented in Section V. Finally, the paper is summarized in Section VI.

II. CLASSIFICATION OF OUTLIER AND DETECTION TECHNIQUES

This section of the research work, describes the different outlier classification and different outlier detection techniques which can efficiently mine the essential and useful information from the outsized health care data.

A. Types of Outlier

An important aspect of an outlier detection technique is the nature of the desired outlier. Outliers can be classified into following three categories:

i). Point Outliers
ii). Contextual Outliers
iii). Collective Outliers.

1. Point Outliers. If an individual data instance can be considered as anomalous with respect to the rest of data, then the instance is termed as a point outlier. This is the simplest type of outlier and is the focus of majority of research on outlier detection. For example, in Figure 1, points o1 and o2 as well as points in region o3 lie outside the boundary of the normal regions, and hence are point outliers since they are different from normal data points. As a real life example, if we consider credit card fraud detection with data set corresponding to an individual's credit card transactions assuming data definition by only one feature: amount spent. A transaction for which the amount spent is very high compared to the normal range of expenditure for that person will be a point outlier.

2. Contextual Outliers. If a data instance is anomalous in a specific context (but not otherwise), then it is termed as contextual outlier (also referred to as conditional outlier). The notion of a context is induced by the structure in the data set and has to be specified as a part of the problem formulation. Each data instance is defined using two sets of attributes [3]:

- Contextual attributes.
  The contextual attributes are used to determine the context (or neighborhood) for that instance. For example, in spatial data sets, the longitude and latitude of a location are the contextual attributes.

In time series data, time is a contextual attribute which determines the position of an instance on the entire sequence.

- Behavioral attributes.
  The behavioral attributes define the non-contextual characteristics of an instance. For example, in a spatial data set describing the average rainfall of the entire world, the amount of rainfall at any location is a behavioral attribute.
  The anomalous behavior is determined using the values for the behavioral attributes within a specific context. A data instance might be a contextual outlier in a given context, but an identical data instance (in terms of behavioral attributes) could be considered normal in a different context. This property is key in identifying contextual and behavioral attributes for a contextual

3. Collective Outliers. If a collection of related data instances is anomalous with respect to the entire data set, it is termed as a collective outlier. The individual data instances in a collective outlier may not be outliers by themselves, but their occurrence together as a collection is anomalous. Figure 2 illustrates an example which shows a human electrocardiogram output [4]. The highlighted region denotes an outlier because the same low value exists for an abnormally long time (corresponding to an Atrial Premature Contraction). It may be noted that low value by itself is not an outlier but its successive occurrence for long time is an outlier.

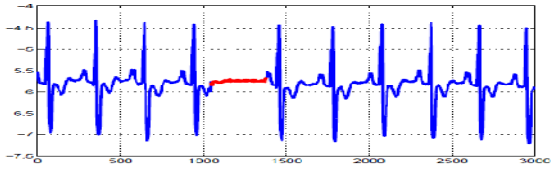

Fig. 2. Collective outlier in a human ECG output corresponding to an Atrial Premature Contraction.

B. Health Care Data Mining Techniques

For the detection of essential and useful information from the outsized health care data, various data mining techniques have been developed and designed, described below:

1. Clustering. The clustering is a common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data [5]. Rui Velosoa [6] had used the vector quantization method in clustering approach in predicting the readmissions in intensive medicine. The algorithms used in vector quantization method are k-means, k-mediods and x-means. The datasets used in this study were collected from patient's clinical process and laboratory results. The evaluations for each of the algorithms are conducted using the Davies-Bouldin Index. The k-means obtained best results while x-means obtained a fair result. The k-mediods obtained the worst results. From results, the work by these researchers provides a useful result in helping to characterize the different types of patients having a higher probability to be readmitted. A more significant comparison on the method cannot be made since this is the only one method in my review discussing on the vector quantization.

2. Classification. Classification comprises of two footsteps: - 1) training and 2) testing. Training builds a classification model on the basis of training data collected for generating classification rules. The IF-THEN prediction rule is highly popular in data mining; they signify facts at a high level of abstraction. The accuracy of classification model hinge on the degree to which classifying rules are true which is estimated by test data [7]. In health care domain, classification can be made useful as "if DiabeticFamilyHistory=yes AND HighSugerIntake=yes THEN DiabetesPossiblity=High". Hatice *et al.,* to analyse skin diseases by using weighted KNN classifier [8].

3. Statistical techniques. The statistical data mining methods effectively consider big data for identifying structures (variables) with the appropriate predictive power in order to yield reliable and robust large-scale statistical models and analyses. The net effect of excessive fraudulent claims is excessive billing amounts, higher per-patient costs, excessive per-doctor patients, higher per-patient tests, and so on. This excess can be identified using special analytical tools. Provider statistics include; total number of patients, total amount billed, total number of patient visits, per-patient average visit numbers, per-patient average billing amounts, per-patient average medical test costs, per-patient average medical tests, per-patient average prescription ratios (of specially monitored drugs) and many more. There are many supervised and unsupervised data mining techniques out of which the following are chosen.

Test. The probabilistic model is not sufficient for detecting outliers. A procedure that determines whether a particular object is an outlier is required. Such a procedure is referred to as a test. A standard test consists in the verification of the basic hypothesis (null hypothesis). The basis hypothesis is a statement that an object fits the probabilistic model of the system, i.e., has been generated by a given distribution law. If there is an alternative hypothesis, such that either the basic or alternative hypothesis is true, the problem of verification of null hypothesis is solved by standard methods of probability theory and mathematical statistics. If there is no alternative hypothesis, the verification is more complicated [9].

SmartSifter. Another interesting approach to detecting outliers by statistical methods is implemented in the Smart- Sifter algorithm [10]. The basic idea of this algorithm is to construct a probabilistic data model based on observations. In this case, only the model, rather than the entire dataset, is stored. The objects are processed successively, and the model learns while processing each data object. A data object is considered to be an outlier if the model changes considerably after processing it. For these purposes, a special metrics, the outlier factor, is introduced to measure changes in the probabilistic model after adding a new element.

Regression Analysis. Methods for detecting outliers based on the regression analysis are also classified among statistical methods. The regression analysis problem consists in finding a dependence of one random variable (or a group of variables) *Y* on another variable (or a group of variables) *X*. Specifically, the problem is formulated as that of examining the conditional probability distribution *Y*|*X*. In the regression methods for the outlier analysis, two approaches are distinguished. In the framework of the first approach, the regression model is constructed with the use of all data; then, the objects with the greatest error are successively, or simultaneously, excluded from the model. This approach is called a reverse search. The second approach consists in constructing a model based on a part of data and, then, adding new objects followed by the reconstruction of the model. Such a method is referred to as a direct search [11].

Then, the model is extended through addition of most appropriate objects, which are the objects with the least deviations from the model constructed. The objects added to the model in the last turn are considered to be outliers. Basic disadvantages of the regression methods are that they greatly depend on the assumption about the error distribution and need a priori partition of variables into independent and dependent ones.

III. PROPOSED METHODOLOGY

Here, we discuss main methodology which we have used in our work namely ID3, k-NN and genetic algorithm (GA).

A. ID3

In decision tree learning, ID3 (Iterative Dichotomiser3) is an algorithm invented by Ross Quinlan [12], used to generate a decision tree from a dataset. ID3 is the precursor to the C4.5 algorithm, and is typically used in the machine learning and natural language processing domains. The ID3 algorithm begins with the original set {\displaystyle S} as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set {\displaystyle S} and calculates the entropy {\displaystyle H(S)} (or information gain {\displaystyle IG(S)}) of that attribute. It then selects the attribute which has smallest entropy (or largest information gain) value. The set {\displaystyle S} is then split by the selected attribute (e.g. age is less than 50, age is between 50 and 100, age is greater than 100) to produce subsets of the data. The algorithm continues to recurse on each subset, considering only attributes never selected before. Recursion on a subset may stop in one of these cases:

- Every element in the subset belongs to the same class (+ or -), then the node is turned into a leaf and labeled with the class of examples.
- There are no more attributes to be selected, but the examples still do not belong to the same class (some are + and some are -), then the node is turned into a leaf and labeled with the most common class of examples in the subset.
- There are no examples in the subset, this happens when no example in the parent set was found to be matching a specific value of the selected attribute, for example if there was no example with age >= 100. Then a leaf is created, and labelled with most common class of examples in the parent set.

Throughout the algorithm, the decision tree is constructed with each non-terminal node representing selected attribute on which the data was split, and terminal nodes representing the class label of the final subset of this branch.

B. K-Nearest Neighbour (k-NN)

In pattern recognition, the k-nearest neighbor algorithm (k-NN) is a non-parametric method used for classification and regression [13]. In both cases, the input consists of k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

The distance to k^{th} nearest neighbor can also be seen as a local density estimate and thus is also a popular outlier score in anomaly detection. The larger the distance to the k-NN, the lower the local density, the more likely and the query point is an outlier [14]. Although quite simple, this outlier model, along with another classic data mining method, local outlier factor, works quite well also in comparison to more recent and more complex approaches, according to a large scale experimental analysis [15].

C. Genetic Algorithm (GA)

Genetic algorithms provide a comprehensive search methodology for machine learning and optimization. Algorithm is started with a set of solutions (represented by chromosomes) called population. Solutions from one population are considered and used to form a new population. This is motivated by a hope, that the new population will be better than the old one. Solutions which are selected to form new solutions (offspring) are selected according to their fitness – the more suitable they are, the more chances they have to reproduce. The genetic algorithm performs the following operations such as crossover, mutation and selection [16].

Crossover. Crossover probability crossover the parents to form a new offspring (children). If no crossover was performed, offspring is an exact copy of parents. We use a Subset Size-Oriented Common Feature Crossover Operator (SSOCF) which keeps useful informative blocks and produces offspring's which have the same distribution than the parents. Offspring's are kept, only if they fit better than the least good individual of the population.

Mutation. During the mutation stage, a chromosome has a probability p_{mut} to mutate. If a chromosome is selected to mutate, we choose randomly a number n of bits to be flipped then n bits are chosen randomly and flipped. The mutation is an operator which allows diversity. With a mutation probability mutate new offspring at each locus (position in chromosome).

Selection. Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected).

D. Proposed Algorithm

Initially, the class-wise probability is settled using eq. (1).

$$\text{Entropy H(X)} = -\sum_{i=1}^{m} pi\ log2\ pi \qquad (1)$$

Then entropy is calculated for each individual attributes. Consequently, the gain is calculated as follows:

$$\text{Gain} = \text{Entropy(X)} - \text{Entropy (X|Y)} \qquad (2)$$

So as per the above process feature reduction has been done, where gain was higher than that attribute has been qualified for the process and less gain was reduced from dataset.

ALGORITHM STEPS:

Step 1: Read an xlsx data from the source and assign into separate variable X_1.

Step 2: Apply an algorithm to individual pattern for test dataset.

Train = knntrain(X_1(train(:,i_1),:),groups(train(:,1))

Step 3: Object with unknown classes found to do with each of the X_1 classifiers predictions.

Step 4: Select the most repeatedly predicted samples.

KNNGA steps:

Step1: Initialize population from X_1.

Step2: Apply genetic search into selected dataset.

Step3: Apply KNN classifier for testing of all goals/data which are classified or not.

Step4: Arranging every attributes as per the predicted their ranks.

Step5: Select higher ranked attribute from among.

Step6: Apply KNN-GA() on each goal subset of the attributes to enhancing the accuracy level.

Step7:If KNN-GA classified correct class then assign into a final predicted class;

Otherwise

data_class = class_knnga; and go to step 8

Step8: Execute the reproduction and apply crossover operator.

Step10: Perform mutation to produce new population X'_1.

Step11: Calculate the local maxima for each category and

Repeat the steps till iteration is not finished

Step12: For each test $X'1$, start all trained base models then prediction of result by combining of all trained models, and separate the misclassified data by optimized

knnGA.

If all data classified and predicted accurate classes then go to step 13;

Otherwise

Go to step step 3 again

Step 13: Classification: goal wise classification result obtained

Step 14: Measurement: Accuracy, sensitivity etc.

Block diagram shown in Fig. 3 depicts working of the proposed approach, where health care dataset is selected for the processing initially, and then entire dataset is logically separate for the moment due to it is containing string fields as well as numeric fields. Hence, in the designing approach we developed separate mechanism for string and numeric data.

Pre-processing**:** It converts the data which is more reliable for unsupervised learning by removing the labels from the dataset.

Data fraction**:** Preprocessed data are used to partition into training & testing sets samples.

If the normal class has been easily detected then its goes to the separately normal class otherwise it will go to the knnGA classifier.

In this process, each class has been accurately predicted with their own identity and after successful prediction the result analysis approach follows for the detected intrusions.

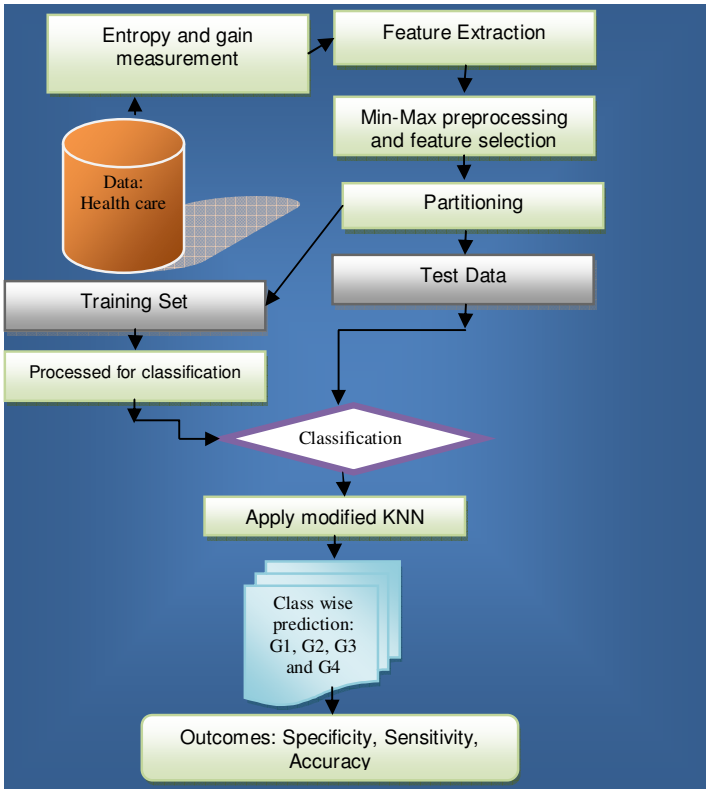

Fig. 3. Block diagram of the proposed methodology.

IV. EXPERIMENTAL RESULTS

In this section, we present the results from our extensive experiments to compare the performance of KNN, local outlier factor (LOF) and PB-KNN algorithms on real health care data from a Chinese city. All the experiments are conducted on MATLAB platform, with three Intel 3.4 GHz machines, each running 16GB RAM.

A. Processed Attributes

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers up to now. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1, 2, 3, 4) from absence (value 0).
Only 14 attributes used:
1. #3 (age) 2. #4 (sex) 3. #9 (cp)
4. #10 (trestbps) 5. #12 (chol) 6. #16 (fbs) 7. #19 (restecg) 8. #32 (thalach) 9. #38 (exang) 10. #40 (oldpeak) 11. #41 (slope) 12. #44 (ca) 13. #51 (thal) 14. #58 (num) (the predicted attribute).

B. GUI Environment

This section shows the main GUI environment in figure 4 of the proposed methodology that contains evolution process of probability entropy gain mechanism in figure 5.

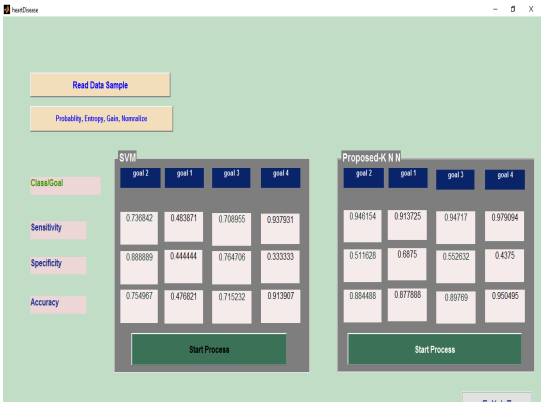

Fig. 4. Main GUI of the proposed methodology.

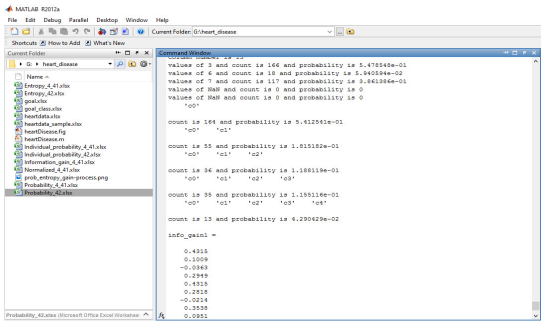

Fig. 5. GUI of probability, entropy gain of the proposed methodology

C. Result Analysis

The result analysis of the proposed work is performed using accuracy and specificity parameter. For the processed attributes confusion matrix is formed, shown in Table 1.

Table 1: Confusion Matrix for SVM.

Confusion Matrix of 'c2'	
98	2
35	16
0	0
Confusion Matrix of 'c1'	
60	15
64	12
0	0
Confusion Matrix of 'c3'	
95	4
39	13
0	0
Confusion Matrix of 'c4'	
136	4
9	2
1	0

Table 2: Confusion Matrix for Proposed Methodology.

Confusion Matrix of 'c2'	
246	21
14	22
0	0
Confusion Matrix of 'c1'	
233	15
22	33
0	0
Confusion Matrix of 'c3'	
251	17
14	21
0	0
Confusion Matrix of 'c4'	
281	9
6	7
0	0

Accuracy Analysis. Accuracy has two definitions:

- More commonly, it is a description of systematic errors, a measure of statistical bias;
- Alternatively, ISO defines accuracy as describing both types of observational error above (preferring the term trueness for the common definition of accuracy).

For this parameter the comparison between SVM and proposed method is perform in which it is found that the accuracy rate of SVM is about 75% and our method is about 90% shown in figure 6 which means proposed method generates better accuracy rate than the existing SVM method.

Table 3: Accuracy result analysis of the proposed method.

Accuracy		
	SVM	**Proposed**
goal2	0.754967	0.88488
goal1	0.476821	0.877888
goal3	0.715232	0.89769
goal4	0.913907	0.950495

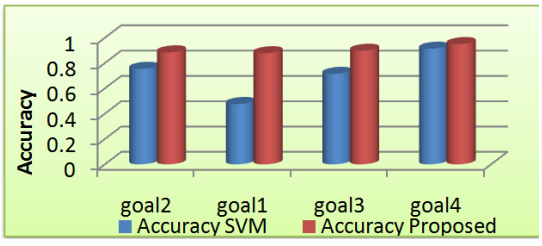

Fig. 6. Accuracy graph between SVM and Proposed method.

Specificity Analysis. Specificity (also called the true negative rate) measures the proportion of negatives that are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition).

For this parameter the comparison between SVM and proposed method is perform in which it is found that the accuracy rate of SVM is about 60% and our method is about 50% shown in figure 7, which means our method generates better specificity rate than the existing SVM method.

Table 4: Specificity result analysis of the proposed method.

Specificity		
	SVM	**Proposed**
goal2	0.888889	0.511628
goal1	0.44444	0.6875
goal3	0.764706	0.552632
goal4	0.33333	0.4375

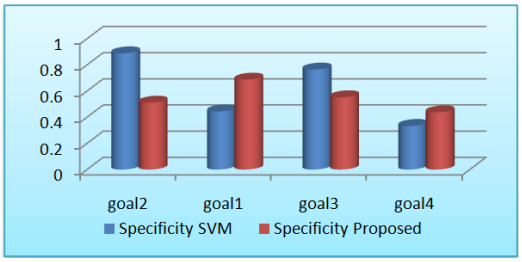

Fig. 7. Accuracy graph between SVM and Proposed method.

V. CONCLUSION AND FUTURE WORK

Due to the extensive growth in digitizing health care data, the storage and security is the major issue and to discover the useful information from the outsized health care data. For outlier detection we propose a hybrid approach which can efficiently detect the outlier in large scale and high dimensional data. The proposed methodology use ID3-KNN&GA for the outlier detection and its simulation is performs in MATLAB simulation toolbox with the performance measuring parameter accuracy and specificity. The result analysis of the accuracy and specificity measures gives improved results than the traditional approach SVM for outlier detection in health care data. In future work, this propose method can be implemented for some another performance measuring parameters of data mining.

REFERENCES

[1]. W. Raghupathi and V. Raghupathi. "Big data analytics in healthcare: promise and potential," *Health Information Science and Systems*, 2014, **2**(1).

[2]. H. C Koh and G. Tan. "Data mining applications in healthcare," *Journal of Healthcare Information Management*, 2005, **19**(2), pp.64-72.

[3]. Song, X., Wu, M., Jermaine, C., and Ranka, S. 2007. Conditional outlier detection. IEEE Transactions on Knowledge and Data Engineering 19, 5, 631-645.

[4]. Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for com-plex physiologic signals. Circulation **101**, 23, e215 - e220. Circulation Electronic Pages: http://circ.ahajournals.org /cgi/content/full/101/23/e215.

[5]. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," AI Mag., pp. 37–54, 1996.

[6]. R. Veloso, F. Portela, M. F. Santos, Á. Silva, F. Rua, A. Abelha, and J. Machado, "A Clustering Approach for Predicting Readmissions in Intensive Medicine," *Procedia Technol.*, vol. **16**, pp. 1307–1316, 2014.

[7]. Jia-Fu Chang & Lei Hua, "Data Mining In Healthcare And Biomedicine: A Survey ff The Literature", Springer, May-2011.

[8]. C. Hattice & K. Metin, "A Diagnostic Software Tool For Skin Diseases With Basic and Weighted k-NN", *Innovations in Intelligent Systems and Applications (INISTA)*, 2012.

[9]. Han, J. and Kamber, M., Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000.

[10]. Yamanishi, K, Takeichi, J., and Williams, G., (2000). On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms, Proc. of the Sixth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, Boston, 2000, pp. 320–324.

[11]. Hadi, A.S., (1992). A New Measure of Overall Potential Influence in Linear Regression, Computational Statistics Data Analysis, 1992, vol. **14**, pp. 1–27.

[12]. Quinlan, J. R. (1986). Induction of Decision Trees. Mach. Learn. **1**, 1 (Mar. 1986), 81-106.

[13]. Altman, N. S. (1992)."An introduction to kernel and nearest-neighbor nonparametric regression". *The American Statistician.* **46**(3): 175–185.

[14]. Ramaswamy, S.; Rastogi, R.; Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. Proceedings of the 2000 ACM SIGMOD international conference on Management of data – SIGMOD ISBN 1-58113-217-4.

[15]. Campos, Guilherme O.; Zimek, Arthur; Sander, Jörg; Campello, Ricardo J. G. B.; Micenková, Barbora; Schubert, Erich; Assent, Ira; Houle, Michael E. (2016). "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study". *Data Mining and Knowledge Discovery*. doi:10.1007/s10618-015-0444-8. ISSN 1384-5810.

[16]. Ranno Agarwal "Genetic Algorithm in Data Mining", *International Journal of Advanced Research in Computer Science and Software Engineering*, **5**(9), September- 2015, pp. 631-634.