IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 3, Ver. I (May.-June. 2017), PP 59-70
www.iosrjournals.org

Review of Medical Disease Symptoms Prediction Using Data Mining Technique

Rahul Deo Sah[1], Dr. Jitendra Sheetalani[2]
[1, 2](Sri Satya Sai University & Medical Sciences, Sehore, India)

Abstract: *Now a day's data mining technique used in the field of medical diagnose of critical diesis and clinical data. The prediction of mining technique is major issue. For the enhancement of mining technique used various approach such as fuzzy logic, feature optimization and machine learning based classification technique. in this classification proceed based on classifier selection to medical disease data and propose a clustering-based classifier selection method. In the method, many clusters are selected for an ensemble process. Then, the standard presentation of each classifier on selected clusters is calculated and the classifier with the best average performance is chosen to classify the given data. In the computation of normal act, weighted average is technique is used. Weight values are calculated according to the distances between the given data and each selected cluster. There are generally two types of multiple classifiers combination: multiple classifiers selection and multiple classifiers fusion. Multiple classifiers selection assumes that each classifier has expertise in some local regions of the feature space and attempts to find which classifier has the highest local accuracy near an unknown test sample. Then, this classifier is nominated to make the final decision of the system.*

Keywords: *Medical Diseases, Data Mining, Classification, Clustering*

I. Introduction

Classification proceed based on classifier selection to medical disease data and propose a clustering-based classifier selection method. In the method, many clusters are selected for a ensemble process. Then, the standard presentation of each classifier on selected clusters is calculated and the classifier with the best average performance is chosen to classify the given data. In the computation of normal act, weighted average is technique is used. Weight values are calculated according to the distances between the given data and each selected cluster. There are generally two types of multiple classifiers combination: multiple classifiers selection and multiple classifiers fusion. Multiple classifiers selection assumes that each classifier has expertise in some local regions of the feature space and attempts to find which classifier has the highest local accuracy in the vicinity of an unknown test sample. Then, this classifier is nominated to make the final decision of the system[23][24][25]. Performance of a classifier is frequently the most important aspect of its value and is measured using a variety of well-known method and matrix is used. On the other hand knowledge of a classifier is often treated as less important or even neglected. However it is vital for the users of the classifier as they belief it more if they can realize how the classifier works and because additional knowledge about the relations in observed data can be extracted by involved classifier. Consequently some of the old methods focus on knowledge of learned classifiers or transforming non- knowledge classifiers into human- knowledge structure[20][22]. There is lack of algorithms that treat accuracy and knowledge of classifiers as uniformly significant, converting the domain of constructing a classifier into heuristic optimization crisis. Such algorithms are especially important in domains where there are parts of attribute space that can be classified with high accuracy using knowledgeable classifier and parts that require non- knowledge classifiers to achieve required classification accuracy[18][21][26].

The process of combining different clustering output (cluster ensemble or clustering Aggregation) emerged as an alternative approach for improving the quality of the Results of clustering algorithms. It is based on the success of the combination of supervised classifiers. Given a set of objects, a cluster ensemble method consists of two principal steps: Generation, which is about the creation of a set of partitions a of these objects, and Consensus Function, where a new partition, which is the integration of all partitions obtained in the generation step, is computed. Over the past years, many clustering ensemble techniques have been proposed, resulting in new ways to face the problem together with new fields of application for these techniques. Besides the presentation of the main methods, the introduction of taxonomy of the different tendencies and critical comparisons among the methods is really important in order to give a practical application to a survey[33][34]. Thus, due to the importance that clustering ensembles have gained facing cluster analysis, we have made a critical study of the different approaches and the existing methods. Feature selection technique is used for selecting subset of relevant features from the data set to build robust classification models. Classification accuracy is improved by removing most irrelevant and redundant features from the dataset. Ensemble model is proposed for improving classification accuracy by combining the prediction of multiple classifiers. In this

dissertation used cluster based ensemble classifier[30][31][32]. The performance of each classifier and ensemble model is evaluated by using statistical measures like accuracy, specificity and sensitivity. Classification of medical data is an important task in the prediction of any disease. It even helps doctors in their diagnosis decisions. Cluster oriented Ensemble classifier is to generate a set of classifiers instead of one classifier for the classification of a new object, hoping that the combination of answers of multiple classification results in better performance. We demonstrate the algorithmic use of the classification technique by extending SVM the most popular binary classification algorithms[27][28][29]. From the studies above, the key to improve cluster oriented classifier is to improve binary classification. In the final part of the thesis, we include empirical evaluation that aim at understanding binary classification better in the context of ensemble learning. Two of the most critical and well generalized problems of medical data are its new evolved feature and concept-drift. Since a medical data is a fast and continuous event, it is assumed to have infinite length. Therefore, it is difficult to store and use all the historical data for training. The most discover alternative is an incremental learning technique. Several incremental learners have been proposed to address this problem [18], [17]. In addition, concept-drift occurs in the stream when the underlying concepts of the stream change over time. A variety of techniques have also been proposed in the literature for addressing concept-drift [12], [16], [17] in data stream classification. However, there are two other significant characteristics of data streams, such as concept evolution and feature evolution that are ignored by most of the existing techniques. Concept-evolution occurs when new classes evolve in the data. Cluster oriented ensemble classifier used to reduce feature evaluation problem in medical disease data classification. The rest of paper describe as in section III. Discuss related work in the field of medical dieses prediction. In section III. Discuss mining technique. in section IV discuss experimental results and finally discuss conclusion & future work in section V.

II. Related Work

In this section discuss the related work in the field of medical science for the prediction of critical disease based on data mining technique. The survey of medical disease used various data mining technique for the prediction [40].

Mai Shouman, Tim Turner and Rob Stocker Et al. [1] This paper distinguishes crevices in the exploration on coronary illness finding and treatment and talked about a model to methodically close those holes to find if applying information mining systems to coronary illness treatment information can give as dependable execution as that accomplished in diagnosing coronary illness. They applying half breed information mining strategies has demonstrated promising outcomes in the conclusion of coronary illness, so applying cross breed information mining procedures in choosing the appropriate treatment for coronary illness patients' needs assist examination. This paper distinguishes holes in the exploration on coronary illness conclusion and treatment and talked about a model to efficiently close those crevices to find if applying information mining procedures to coronary illness treatment information can give as solid execution as that accomplished in diagnosing coronary illness patients. M. L. Kowalski, J. S. Makowska, M. Blanca, S. Bavbek, G. Bochenek, J. Bousquet, P. Bousquet, G. Celik, P. Demoly, E. R. Gomes, E. Ni_zankowska-Mogilnicka, A. Romano, M. Sanchez-Borges, M. Sanz, M. J. Torres, A. De Weck, A. Szczeklik and K. Brockow Et al. [2] as indicated by specialists, Nonsteroidal hostile to inflammatory drugs (NSAIDs) are in charge of 21–25% of detailed unfriendly medication occasions which incorporate immunological and nonimmunological excessive touchiness responses. This review displays state-of-the-art data on patho-systems, clinical range, symptomatic devices and administration of excessive touchiness responses to NSAIDs. Clinically, NSAID extreme touchiness is especially showed by bronchial asthma, rhinosinusitis, hypersensitivity or urticaria and assortment generally cutaneous and organ-specific responses. Analysis of excessive touchiness to a NSAID incorporates comprehension of the fundamental instrument and is essential for anticipation and administration. A stepwise way to deal with the analysis of excessive touchiness to NSAIDs is examined, including clinical history, in vitro testing or potentially incitement test with a guilty party or option sedate contingent upon the kind of the response. The analytic procedure ought to bring about furnishing the patient with composed data both on illegal and on option drugs. Chaitrali S. Dangare and Sulabha S. Apte Et al. [3] This paper has dissected forecast frameworks for Heart illness utilizing more number of information traits. The framework utilizes medicinal terms, for example, sex, circulatory strain, cholesterol like 13 ascribes to foresee the probability of patient getting a Heart ailment. From results, it has been seen that Neural Networks gives exact outcomes as contrast with Decision trees and Naive Bayes. This framework can be further extended. It can utilize more number of information qualities recorded above in table 1 and 2. Other information mining strategies can likewise be utilized for predication e.g. Grouping, Time arrangement, Association rules. The content mining can be utilized to mine colossal measure of unstructured information accessible in human services industry database. Asha Rajkumar and Mrs. G.Sophia Reena Et al. [4] In this paper the information characterization depends on regulated machine learning calculations which result in precision, time taken to construct the calculation. Tanagra instrument is utilized to group the information and the information is assessed utilizing 10-overlap

cross approval and the outcomes are thought about. This paper manages the outcomes in the field of information arrangement got with Naive Bayes calculation, Decision list calculation and KNN calculation, and in general execution made known Naive Bayes Algorithm when tried on coronary illness datasets. The time taken to run the information for result is quick when contrasted with different calculations. It demonstrates the improved execution as per its quality. Traits are completely characterized by this calculation and it gives 52.33% of exact outcome. In light of the trial comes about the grouping precision is observed to be better utilizing Naive Bayes calculation contrast with different calculations. From the above outcomes, Naive Bayes calculation assumes a key part in molding enhanced characterization precision of a dataset. Nidhi Bhatla and Kiran Jyoti Et al. [5] This paper goes for investigating the different information mining methods presented as of late for coronary illness forecast. The perceptions uncover that Neural systems with 15 qualities has beated over every other dat mining procedures. Another conclusion from the examination is that choice tree has additionally demonstrated great exactness with the assistance of hereditary calculation and highlight subset determination. Different methods and information mining classifiers are characterized in this work which has risen lately for proficient and viable coronary illness determination. The examination demonstrates that Neural Network with 15 properties has demonstrated the most astounding precision i.e. 100% up until this point. Then again, Decision Tree has likewise performed well with 99.62% precision by utilizing 15 traits. Additionally, in blend with Genetic Algorithm and 6 qualities, Decision Tree has indicated 99.2% proficiency. Shweta Kharya Et al. [6] In this paper This review paper condenses different audit and specialized articles on bosom growth conclusion and guess likewise they concentrate on ebb and flow research being done utilizing the information mining procedures to upgrade the bosom malignancy analysis and anticipation. Different information mining strategies have been broadly utilized for bosom growth conclusion. In this paper, they have talked about some of compelling strategies that can be utilized for bosom disease grouping. Among the different information mining classifiers and delicate registering approaches, Decision tree is observed to be best indicator with 93.62% Accuracy on benchmark dataset (UCI machine learning dataset) and furthermore on SEER dataset. HianChye Koh and Gerald Tan Et al. [7] This article investigates information mining applications in medicinal services. Specifically, it examines information mining and its applications inside human services in significant ranges, for example, the assessment of treatment adequacy, administration of social insurance, client relationship administration, and the identification of extortion and manhandle. It likewise gives an illustrative case of a medicinal services information mining application including the recognizable proof of hazard variables related with the onset of diabetes. At long last, the article highlights the impediments of information mining and examines some future headings. Information digging undertakings can fall flat for an assortment of reasons, for example, absence of administration support, impossible client desires, poor venture administration, deficient information mining skill, and the sky is the limit from there. Information mining requires serious arranging and techno-coherent readiness work. M. Anbarasi, e. Anupriya and n.ch.s.n.iyengar Et al. [8] In their work, Genetic calculation is utilized to decide the properties which contribute more towards the analysis of heart diseases which by implication lessens the quantity of tests which are should have been taken by a patient. Thirteen ascribes are decreased to 6 qualities utilizing hereditary inquiry. The target of their work is to anticipate all the more precisely the nearness of coronary illness with decreased number of properties. Innocent Bayes performs reliably previously, then after the fact lessening of qualities with a similar model development time. Order through grouping performs poor contrasted with other two strategies. Irregularities and missing qualities were settled before model development yet continuously, that is not the situation. they expect to broaden their work applying fluffy learning models to assess the force of cardiovascular malady. Sri Harsha Vege Et al. [9] In this postulation they examined an outfit approach for highlight choice, where different element determination strategies are joined to yield more strong and stable outcomes. Troupe of numerous element positioning systems is performed in two stages. The initial step includes making an arrangement of various element selectors, each giving its sorted request of elements, while the second step totals the aftereffects of all component positioning systems. The outfit strategy utilized as a part of their review is recurrence number which is joined by intend to determine any recurrence tally impact. To determine the issue of recurrence crash, they have presented mean requesting. The investigations were directed on two biomedical datasets. The outcomes exhibited that the outfit method performed preferred by and large over any individual ranker. The outcomes additionally demonstrated that the exhibitions of arrangement models are enhanced even after 90% of the components are expelled. Dr. R. GeethaRamani and G. Sivagami Et al. [10] This paper means to give a review of ebb and flow strategies of learning disclosure in databases utilizing information mining systems that are being used today for the order of Parkinson Disease. The paper is expected to check the viability of the utilization of different classifiers to the Parkinson Dataset. This dataset contains 22 properties with different scope of qualities. A relative investigation of a few calculations on the dataset is performed. This is finished by first doing the component importance on the dataset. At that point, the classifiers are actualized upon the dataset. Early recognition of any sort of illness is a basic component. This aides in treating the patient well ahead. In this exploration paper, Random Tree classifier yields the 100% precision.Thamilselvan P and Dr. J. G. R. Sathiaseelan Et al. [11] Image order is an

imperative procedure to create profitable data. The arrangement technique gives the precise outcome in their objective class. This survey thinks about some overwhelming half breed grouping calculations to discover the characterization precision for different informational indexes and their execution of methods. It gives some vital cross breed methods that have been utilized for picture arrangement. In this paper, the crossover information mining calculations are concentrated like GA-SVM, EKM-EELM, AdaBoost-SVM, Decision Tree-Naive Bayes, and SVM-CART. Picture Classification is the most significant piece of picture mining and advanced picture handling. They have pondered development of characterization strategy in view of various half and half approach. The execution of half and half strategies was examined in view of the order precision, preferences and attributes of the reference information focuses. From this review, the half and half strategy NB-SVM demonstrates better exactness in picture arrangement when contrasted with other cross breed approaches. RatnadipAdhikari and R.K. Agrawal Et al. [12] In this paper, they talked about a nonlinear weighted group component for consolidating conjectures from different time arrangement models. they examined a novel nonlinear weighted troupe procedure for estimates blend. It is an expansion of the regular direct blend plot with a specific end goal to incorporate conceivable relationship effects between the taking an interest gauges. An efficient progressive approval system is recommended for deciding the suitable blend weights. The experimental outcomes with three true time arrangement and three determining strategies exhibit that their talked about procedure significantly beats every individual technique as far as got figure correctness's. In addition, it likewise gives significantly preferred outcomes over the great straightforward normal joining procedure. S. D. Kotal and S. K. Roy Bhowmik Et al. [13] The strategy is produced utilizing various straight relapse procedure with five part models, to be specific ECMWF, NCEP, MM5, QLM and JMA. The model parameters are chosen in view of the specimen database of cyclonic frameworks that happened in 2008 and pre-storm season in 2009. The execution of the model is tried utilizing the autonomous specimens that happened amid 2010 and post-storm season in 2009. ECMWF model is observed to be the best among part models. The consequences of this review utilizing the information of 2008 to 2009 are extremely encouraging. they in-have a tendency to incorporate the all information of tornados amid 2008 to 2010 for relapse of MME method for determining the violent wind track of 2011. An aggregate inclination remedy is incorporated into the gathering system as a different direct relapse based minimization guideline for the model figure position against to the watched position is connected in this review. they additionally mean to develop the work of individual inclination evacuation of part models. S. Kotsiantis, K. Patriarcheas and M. Xenos Et al. [14] This paper means to fill the hole between experimental expectation of understudy execution and the current ML methods in a separation instruction condition. The examined method, expects to explore wonders in instructive technique from the purpose of causal understanding perspective. This approach is regularly unfeasible both as far as space and refresh time for online settings with asset imperatives. To help ease the space issue, the extent of the dataset by just putting away and using the latest or most essential occasions could be restricted. This exactness comes to the 73% in the underlying estimates, which depend on statistic information of the understudies and beats the 82% preceding the final examinations. Dataset is from the module 'Presentation in informatics' however a large portion of the conclusions are colossal and show enthusiasm for the dominant part of HOU modules. It is intriguing to contrast their outcomes and those from other open and separation learning programs offered by other open Universities. Given that the HOU is not an ordinary college. Mehdi HosseinzadehAghdam and PeymanKabiri Et al. [15] particle is an essential theme in data security. The motivation behind this review is to distinguish vital components in building an interruption location framework with the end goal that they are computationally efficient and effective. To enhance the execution of interruption discovery framework, this paper talked about an interruption identification framework that its elements are ideally chosen utilizing subterranean insect province advancement. The talked about technique is effectively actualized and has a low computational unpredictability because of utilization of a simplified include set for the classification. Tests and examinations are performed on KDD Cup 99 and NSL-KDD informational collections, the test sets contain 17 sorts of different assaults. The talked about technique decreased the quantity of components by roughly 88% and the recognition blunder lessened by around 24% utilizing KDD Cup 99 test informational collection. This demonstrates the examined strategy is exceptionally solid for interruption location. Comes about show that the examined ACO-based recognition strategy beats different strategies since it can give better and more powerful portrayal of the information. This is because of the way that it can precisely recognize a more extensive scope of assaults utilizing more modest number of components.

III. Data Mining Technique

Knn Classifier

Nearest neighbor classifiers are based on learning by analogy. The training samples are described by n dimensional numeric attributes. Each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample.

"Closeness" is defined in terms of Euclidean distance, where the Euclidean distance, where the Euclidean distance between two points, X=(x1,x2,……,xn) and Y=(y1,y2,….,yn) is dist (X1,X2)= The unknown sample is assigned the most common class among its k nearest neighbors. When k=1, the unknown sample is assigned the class of the training sample that is closest to it in pattern space. Nearest neighbor classifiers are instance-based or lazy learners in that they store all of the training samples and do not build a classifier until a new(unlabeled) sample needs to be classified. This contrasts with eager learning methods, such a decision tree induction and back propagation, which construct a generalization model before receiving new samples to classify. Lazy learners can incur expensive computational costs when the number of potential neighbors (i.e., stored training samples) with which to compare a given unlabeled sample is great. Therefore, they require efficient indexing techniques. An expected lazy learning methods are faster ate training than eager methods, but slower at classification since all computation is delayed to that time. Unlike decision tree induction and back propagation, nearest neighbor classifiers assign equal weight to each attribute. This may cause confusion when there are many irrelevant attributes in the data. Nearest neighbor classifiers can also be used for prediction, that is, to return a real-valued prediction for a given unknown sample. In this case, the classifier returns the average value of the real-valued associated with the k nearest neighbors of the unknown sample[23][36]. The k-nearest neighbors' algorithm is amongst the simplest of all machine learning algorithms. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors. K is a positive integer, typically small. If k = 1, then the object is simply assigned to the class of its nearest neighbor. In binary (two class) classification problems, it is helpful to choose k to be an odd number as this avoids tied votes. The same method can be used for regression, by simply assigning the property value for the object to be the average of the values of its k nearest neighbors. It can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones[39]. The neighbors are taken from a set of objects for which the correct classification (or, in the case of regression, the value of the property) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. In order to identify neighbors, the objects are represented by position vectors in a multidimensional feature space. It is usual to use the Euclidian distance, though other distance measures, such as the Manhanttan distance could in principle be used instead. The k-nearest neighbor algorithm is sensitive to the local structure of the data.

Rough Set Theory

The rough sets theory was proposed by Pawlak in 1982 to deal with uncertain and fuzzy materials and to simplify knowledge. In the rough sets theory, humans use their general knowledge to classify the world around them as abstract or concrete. Everything is classified according to its characteristics, and those with nearly identical characteristics may be put into the same group. This is called indiscernible relation, denoted as Ind and is the basis of rough sets theory. One of the main advantages of rough set theory is that it does not need any preliminary or additional information about data. The main problems that can be approached using rough sets theory include data reduction, discovery of data dependencies, and estimation of data significance, generation of decision algorithms from data, approximate classification of data, discovery of patterns in data and discovery of cause-effect relationships [10]. The following is the concept of rough sets theory.

Knowledge can be finished by the information systems; the basic composition of an information system is the set of objects which are to be studied. The knowledge of these objects is described by their attributes and attributes values. The information system is defined as follows[37]:

$$IS = (U, A) \quad (1)$$

Where U is the universe, a finite non-empty set of objects, $U = \{x_1, x_2, ..., x_m\}$, and A is the set of attributes. Each attribute $a \in A$ (attribute a belonging to the considered set of attributes A) defines an information function:

$$f_a : U \to V_a \quad (2)$$

where V_a is the set of values of a, called the domain of attribute a. In all attributes, there are decision attributes and condition attributes.

Indiscernible relation

For every set of attributes $B \subset A$, an indiscernible relation $Ind(B)$ is defined in the following way: two objects, x_i and x_j, are indiscernible by the set of attributes B in A, if $b(x_i) = b(x_j)$ for every $b \subset B$. The equivalence class of $Ind(B)$ is called the elementary set in B because it represents the smallest discernible groups of objects. For any element x_i of A, the equivalence class of x_i in relation $Ind(B)$ is represented as $[x_i]_{Ind(B)}$.

The rough sets approach to data analysis hinges on two basic concepts, namely the lower and the upper approximations of a set, referring to: the elements that doubtlessly belong to the set, and the elements that possibly belong to the set. The definition is shown as follows:

Let X denote the subset of elements of the universe U, the lower approximation of X in B, denoted as $\underline{BX}$, and is defined as the union of all these elementary sets which are contained in X. More formally[38]:

$$\underline{BX} = \{x_i \in U \left| [x_i]_{Ind(B)} \subset X \right.\} \quad (3)$$

The above statement is to be read as: the lower approximation of the set X is a set of objects x_i , which belong to the elementary sets contained in X (in the space B), $\underline{BX}$ is called the lower approximation of the set X in B.

The upper approximation of the set X, denoted as $\overline{BX}$, is the union of these elementary sets, which have a non-empty intersection with X:

$$\overline{BX} = \{x_i \in U \left| [x_i]_{Ind(B)} \cap X \neq 0 \right.\} \quad (4)$$

The above statement is to be read as: the upper approximation of the set X is a set of objects x_i , which belong to the elementary sets that have a non-empty intersection with X, $\overline{BX}$ is called the upper approximation of the set X in B.

The difference is called a boundary of X in U.

$$BNX = \overline{BX} - \underline{BX} \quad (5)$$

The concepts of core and reduct are two very important concepts of the rough sets theory. If the set of attributes is dependent, one can be interested in finding all possible minimal subsets of attributes. These lead to the same number of elementary sets as the whole set of attributes (reducts) in finding the set of all indispensable attributes (core). Simplification of the information system can be used to recognize some values of attributes which are not necessary for the system[39]. For example, some attributes which are redundant can be deleted or be filtered by means of the simplification procedures. If, $Ind(A) = Ind(A - a_i)$, then the attribute a_i is dispensable, otherwise, a_i is indispensable in A. In other words, if after deleting the attribute a_i , the number of elementary sets in the information system is the same, then it concludes that attribute a_i is dispensable. Hence, the simplification can contain the minimal subsets of independent attributes, which ensure they can represent the whole set. The core is the necessary element for representing knowledge or rules, and is the common part of all reducts. The researcher uses the discernibility matrix to compute the values of reducts and core.

Svm (Support Vector Machine)

Support vector Machine is binary classifier, the performance of classification of support vector machine is high in compression of another binary classifier such as decision tree, KNN and bay,s classifier[35]. Support Vector Machine (SVM) is a novel machine learning method based on statistical learning theory developed by V.N.Vapnik, and it has been successfully applied to numerous classification and pattern recognition problems such as text categorization, image recognition and bioinformatics. It is still in the development stage now.SVM can be used for pattern recognition, regression analysis and principal component analysis. The achievements of SVM in training have Platt's the sequential minimal optimization method, Osuna's the method of Chunking, Joachims' SVM light method and so on. These methods are directed at the training process, and not related to classification process. In the process of SVM training, all the samples are used. So it has no effect on the speed of the classification. Lee and others propose a method of reduction SVM training time and adding the speed of training, reduced support vector machines. The method in the training process is not used in all the samples but by randomly selecting one of the subsets to train, which is through reducing the scale of training to achieve the objective of speeding up the training pace. At the same time, because of the reduction of the support vector quantity, the speed of classification is improved to some degree. However, due to the loss of some support vector classification, precision has declined, especially when the number of support vector is so many that the accuracy of its classification will decline. Burges put forward a way of increasing the speed of Classification, which does not use the support vector in the category function but use a reduction of vector set, which is different from the standard vector set .That is neither training samples nor support vector but it is the transformation of the special vector. The method achieved certain results, but in the process of looking for the reduction of the vector collection, the cost of calculation paid is too large to widely use in practice. The concept of SVM is to transform the input vectors to a higher dimensional space Z by a nonlinear transform, and then an optical hyperplane which separates the data can be found. This hyperplane should have the best generalization capability. As shown in Figure 4.1, the black dots and the white dots are the training dataset which belong to two classes. The Plane H series are the hyperplanes to separate the two classes.

The optical plane H is found by maximizing the margin value $2/\|w\|$. Hyperplanes H_1 and H_2 are the planes on the border of each class and also parallel to the optical hyperplane H. The data located on H_1 and H_2 are called support vectors.

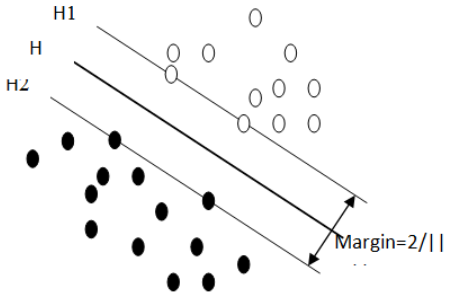

Figure 1 The SVM binary classifications

For training data set $(x_1, y_1),...,(x_l, y_l), y_i \in \{-1,1\}$, to find the optical hyperplane H, a nonlinear transform, $Z = \Phi(x)$, is applied to x, to make x become linearly dividable. A weight w and offset b satisfying the following criteria will be found:

$$\begin{cases} w^T z_i + b \geq 1, & y_i = 1 \\ w^T z_i + b \leq -1, & y_i = -1 \end{cases} \quad \text{.....(a1)}$$

i.e.

$$y_i(w^T z_i + b) \geq 1, \quad i = 1,2,...,l \quad \text{.....(a1.1)}$$

Assume that the equation of the optical hyperplane H (Fig.4.1) is $w_0{}^T z + b_0 = 0$, then the distance of the data point in any of the two classes to the hyperplane is:

$$\rho(w,b) = \min_{x|y=1} \frac{z^T w}{\|w\|} - \max_{x|y=-1} \frac{z^T w}{\|w\|} \quad \text{.....(a2)}$$

A w_0 is to be found to maximize

$$\rho(w_0, b_0) = 2/\|w_0\| = 2/\sqrt{w_0{}^T w_0} \quad \text{.....(a3)}$$

Then the search of the optimal plane H turns to a problem of a second order planning problem.

$$\min_{w,b} \Phi(w) = \frac{1}{2}(w^T w) \quad \text{.....(a4)}$$

Subject to $y_i(w^T z_i + b) \geq 1, \quad i = 1,2,...,l$ .....(a5)

If the sample data is not linearly dividable, find the minimum value of

$$\Phi(w) = \frac{1}{2} w^T w + C\sum_{i=1}^{l} \xi_i \quad \text{.....(a6)}$$

Whereas ξ can be understood as the error of the classification and C is the penalty parameter for this term. By using Lagrange method, the decision function of

$$w_0 = \sum_{i=1}^{l} \lambda_i y_i z_i \quad \text{.....(a7)}$$

will be

$$f = \text{sgn}[\sum_{i=0}^{l} \lambda_i y_i (z^T z_i) + b] \quad \text{.....(a8)}$$

From the functional theory, a non-negative symmetrical function $K(u,v)$ uniquely define a Hilbert space H, K is the rebuild kernel in the space H:

$$K(u,v) = \sum_i \alpha \varphi_i(u) \varphi_i(v) \quad \text{.....(a9)}$$

This stands for an internal product of a characteristic space:

$$z_i^T z = \Phi(x_i)^T \Phi(x) = K(x_i, x) \quad \text{.....(a10)}$$

Then the decision function can be written as:

$$f = \text{sgn}[\sum_{i=1}^{l} \lambda_i y_i K(x_i, x) + b] \quad \text{.....(a11)}$$

The development of a SVM image classification model depends on the selection of kernel function K. There are several kernels that can be used in Support Vector Machines models. These include linear, polynomial, radial basis function (RBF) and sigmoid function:

$$K(x_i, x_j) = \begin{cases} x_i^T x_j & Linear \\ (\gamma x_i^T x_j + coefficient)^{degree} & Polynomial \\ \exp(-\gamma \mid x_i - x_j \mid^2 & RBF \\ \tanh(\gamma x_i^T x_j + coefficient) & Sigmoid \end{cases} \quad \text{.....(a12)}$$

The RBF is by far the most popular choice of kernel types used in Support Vector Machines. This is mainly because of their localized and finite responses across the entire range of the real x-axis.

Improper kernel function might generate poor performance. Currently there is no effective "learning" method to choose a proper kernel function for a specific problem. The selection is decided by the experiment result at this time. In our proposed system, two kernel functions are tested: Radial Basis Function-RBF and Polynomial Function.

$$K_{poly}(x_1, x_2) = (x_1 * x_2 + 1)^p \quad \text{.....(a13)}$$

$$K_{RBF}(x_1, x_2) = \exp(-p \parallel x_1 - x_2 \parallel^2) \quad \text{.....(a14)}$$

IV. Experimental Analysis

In this section, we perform experimental process of CBA, RGI,KNN and SVM algorithm. The methods implement in MATLAB 7.8.0 and tested with very reputed data set from UCI machine learning research center. In the research work, I have measured classification, Accuracy and execution time of ensemble method. To evaluate these performance parameters, I have used five datasets from UCI machine learning repository [10] namely blood dataset, diabetes dataset, Hagerman dataset, heart dataset and liver dataset.

Table 1: shows that the Accuracy and Elapsed time with using CBA and RGI techniques for the same and different dataset.

DATASET NAME	METHOD	SUPPORT	CONFIDENCE	ACCURACY (%)	ELAPSED TIME (SEC)
Liver	**CBA**	0.3	0.5	81.00	6.245
	RGI	0.3	0.5	82.00	5.308
Diabetes	**CBA**	0.3	0.5	81.49	7.245
	RGI	0.3	0.5	83.32	8.451

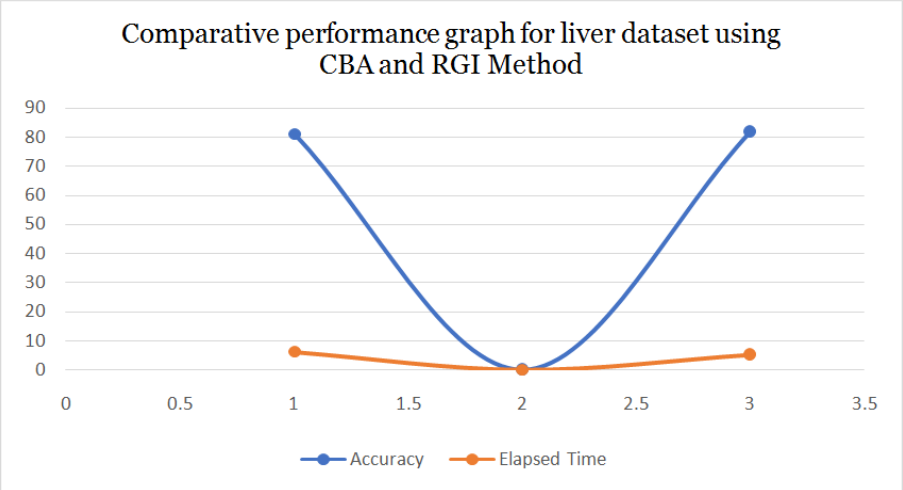

Fig. 2: Shows that comparative result of Liver data set, with using CBA and RGI method.

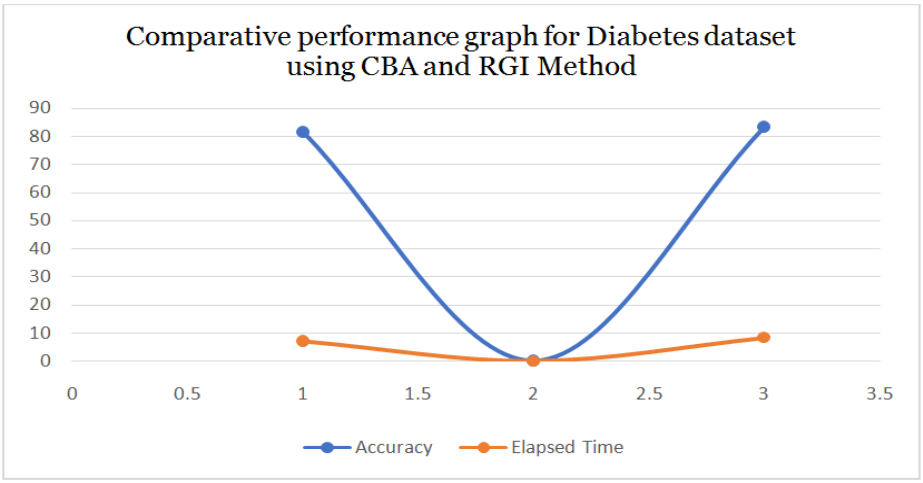

Fig.3: Shows that comparative result of Diabetes data set, with using CBA and RGI method.

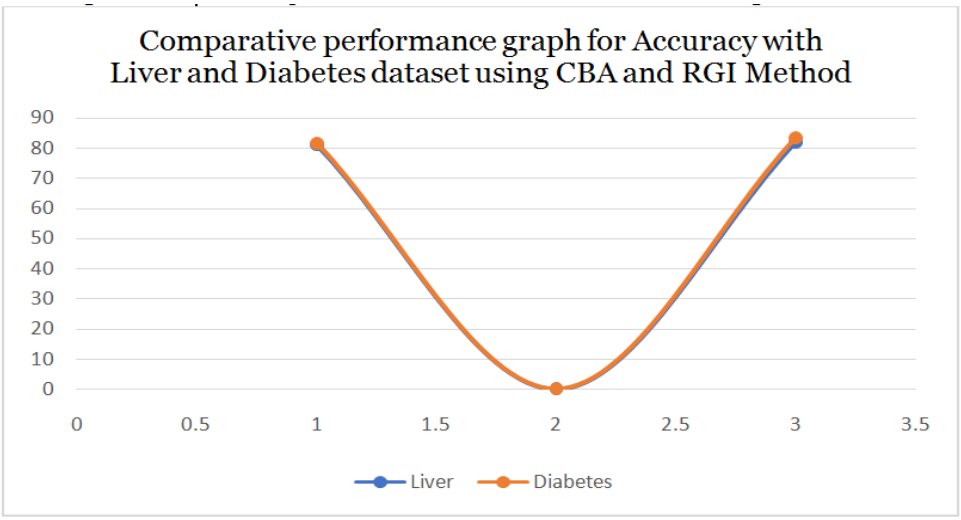

Fig. 4: Shows that comparative result of Accuracy for Liver and Diabetes data set, with using CBA and RGI method.

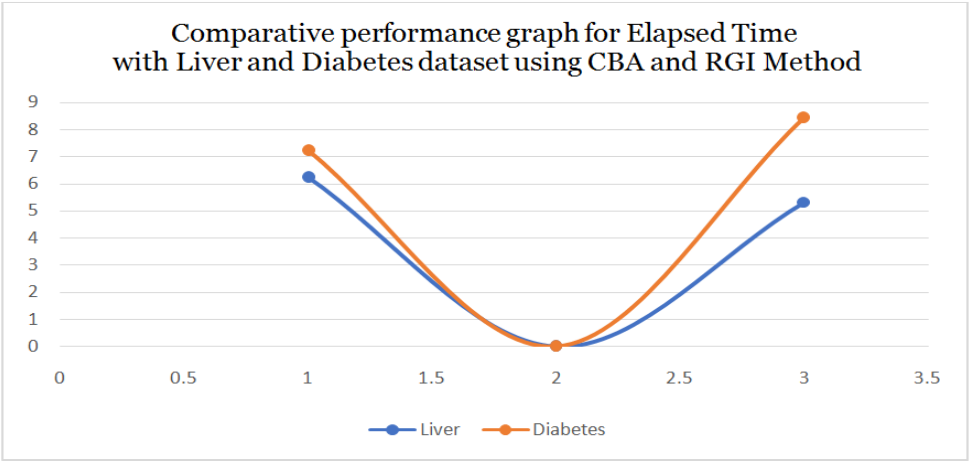

Fig. 5: Shows that comparative result of Elapsed Time for Liver and Diabetes data set, with using CBA and RGI method.

Table 2: Comparative performance evaluation for classification using KNN, SVM and SVM-ANT classifier.

Method Name	Elapsed Time	Mean Absolute Error	Mean Relative Error	Accuracy
KNN	24.43	41.00	26.81	48.18
SVM	24.46	39.50	24.23	59.18

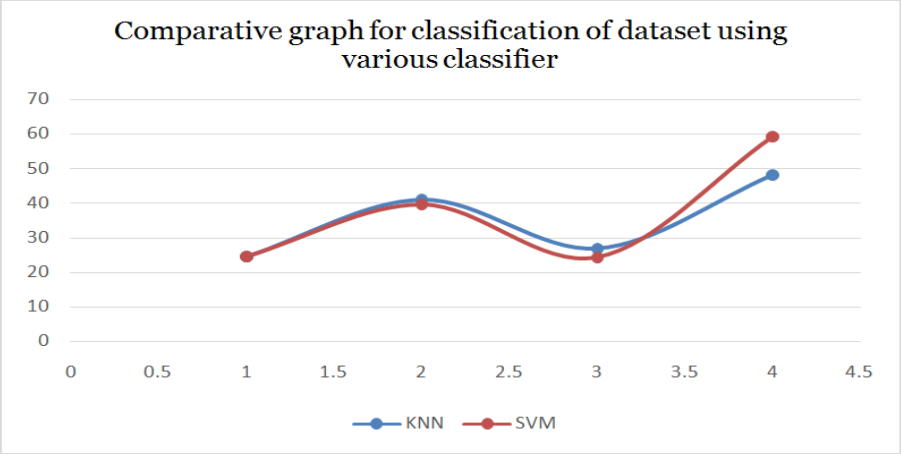

Fig. 6: Comparative performance evaluation for classification using KNN and SVM classifier, here we find the value of elapsed time, mean absolute error, mean relative error and Accuracy.

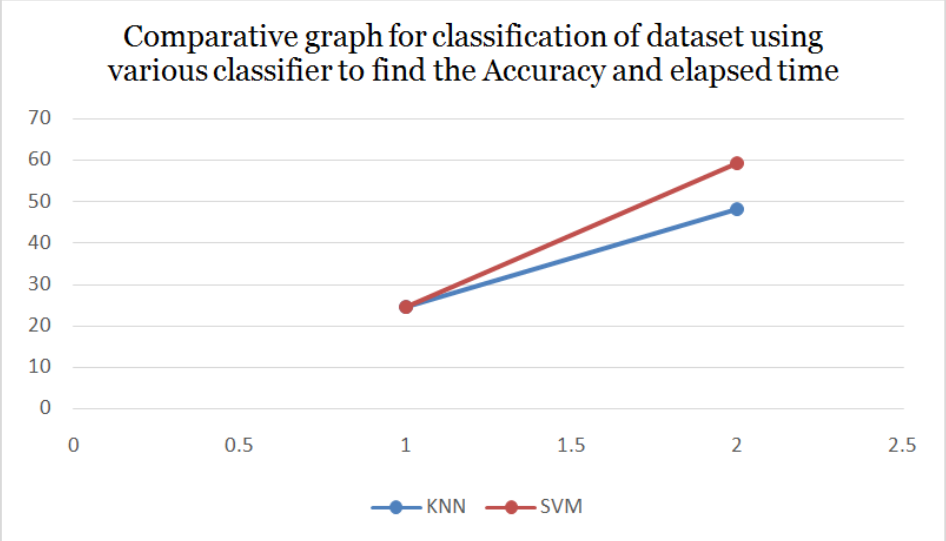

Fig. 7: Comparative performance evaluation for classification using KNN and SVM classifier, here we find the value of elapsed time and Accuracy.

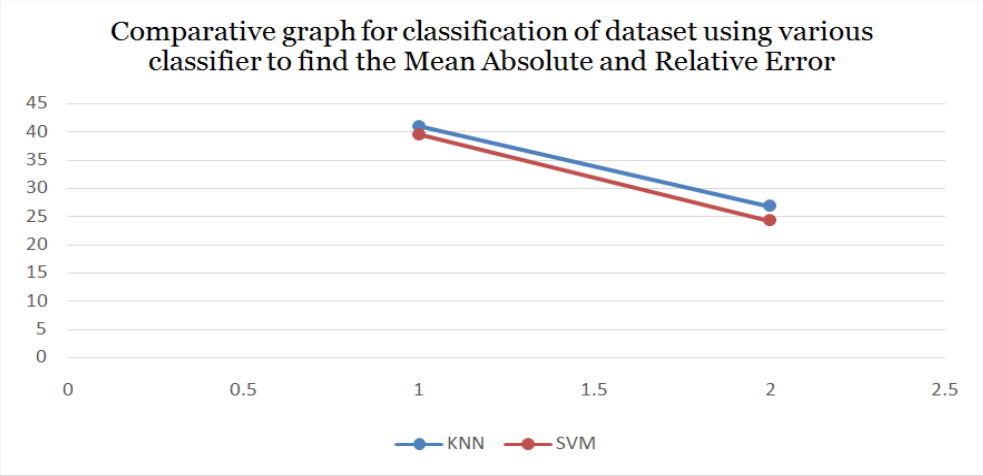

Fig 8: Comparative performance evaluation for classification using KNN and SVM classifier, here we find the value of Mean absolute error and mean relative error.

V. Conclusion & Future Scope

In this paper evaluate the performance of medical disease prediction based on data mining technique. The classifier classified the medical diagnosis of disease data such as cancer, liver problem, and heart disease and so on. SVM method better classified data in compression of conventional cluster ensemble technique. For the classification purpose used two base classifier KNN and SVM. The task of clustering performs by fixed clustering technique such as k-means algorithm. Compared classification accuracy on the same datasets between conventional and proposed ensemble classification algorithms. We investigated the classification accuracy on benchmark datasets including UCI repository, medical disease data, as well as real world datasets. The average and the maximum classification accuracy computed by conventional and proposed ensembles on the same datasets were compared. The comparison results prove that the average classification accuracy and the maximum classification accuracy resulted by proposed ensembles on more datasets is better than the average. In near future also focus on optimized selection of number of base classifier for proper selection of classifier in ensemble process. The diversity of medical diagnosis of disease data are increase day to day. Now in future dimension reduction process are also involved in ensemble classification technique.

References

[1] Mai Shouman, Tim Turner and Rob Stocker, using data mining techniques in heart disease diagnosis and treatment, IEEE, 2012, 189-193.

[2] M. L. Kowalski, J. S. Makowska, M. Blanca, S. Bavbek, G. Bochenek, J. Bousquet, P. Bousquet, G. Celik, P. Demoly, E. R. Gomes, E. Ni_zankowska-Mogilnicka, A. **Romano**, M. Sanchez-Borges, M. Sanz, M. J. Torres, A. De Weck, A. Szczeklik and K. Brockow, Hypersensitivity to nonsteroidal anti-inflammatory drugs (NSAIDs) – classification, diagnosis and management: review of the EAACI/ENDA and GA2LEN/HANNA, John Wiley & Sons A/S, 2011, 818-829.

[3] Chaitrali S. Dangare and Sulabha S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications, 2012, 44-48.

[4] Asha Rajkumar and Mrs. G.Sophia Reena, Diagnosis Of Heart Disease Using Datamining Algorithm, Global Journal of Computer Science and Technology, 2010, 38-43.

[5] Nidhi Bhatla and Kiran Jyoti, An Analysis of Heart Disease Prediction using Different Data Mining Techniques, International Journal of Engineering Research & Technology, 2012, 1-4.

[6] Shweta Kharya, using data mining techniques for diagnosis and prognosis of cancer disease, IJCSEIT, 2012, 55-66.

[7] HianChye Koh and Gerald Tan, Data Mining Applications in Healthcare, Journal of Healthcare Information Management, 2011, 64-71.

[8] M. ANBARASI, E. ANUPRIYA and N.CH.S.N. IYENGAR, Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm, International Journal of Engineering Science and Technology, 2010, 5370-5376.

[9] Sri Harsha Vege, Ensemble of Feature Selection Techniques for High Dimensional Data, Masters Theses & Specialist Projects, 2012, 1-44.

[10] Dr. R. GeethaRamani and G. Sivagami, Parkinson Disease Classification using Data Mining Algorithms, International Journal of Computer Applications, 2011, 17-22.

[11] Thamilselvan P and Dr. J. G. R. Sathiaseelan, Image Classification using Hybrid Data Mining Algorithms – A Review, IEEE, 2015, 1-6.

[12] RatnadipAdhikari and R.K. Agrawal, A Novel Weighted Ensemble Technique for Time Series Forecasting, Springer, 2012, 38-49.

[13] S. D. Kotal and S. K. Roy Bhowmik, A multimodel ensemble (MME) technique for cyclone track prediction over the North Indian Sea, GEOFIZIKA, 2011, 275-291.

[14] S. Kotsiantis, K. Patriarcheas and M. Xenos, A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education, Knowledge-Based Systems, 2010, 529–535.

[15] Mehdi HosseinzadehAghdam and PeymanKabiri, Feature Selection for Intrusion Detection System Using Ant Colony Optimization, International Journal of Network Security, 2016, 420-432.

[16] Sarwesh Site and Dr. Sadhna K. Mishra, A Review of Ensemble Technique for Improving Majority Voting for Classifier, International Journal of Advanced Research in Computer Science and Software Engineering, 2013, 177-180.

[17] SANDRO VEGA-PONS and JOSÉ RUIZ-SHULCLOPER, A SURVEY OF CLUSTERING ENSEMBLE ALGORITHMS, International Journal of Pattern Recognition and Artificial Intelligence, 2011, 337-372.

[18] Zenon Brzoza, Canonica Walter, Martin K Church and Martin Metz, The EAACI/GA2LEN/EDF/WAO Guideline for the definition, classification, diagnosis, and management of urticaria: the 2013 revision and update, John Wiley & Sons A/S. Published by John Wiley & Sons Ltd, 2014, 1-20.

[19] ROMAIN A. PAUWELS, A. SONIA BUIST, PETER M. A. CALVERLEY, CHRISTINE R. JENKINS and SUZANNE S. HURD, Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease, Am J RespirCrit Care Med., 2011, 1256-1276.

[20] Arie Levine, James Markowitz, Mary E Sherlock and Jeffrey Hyams, Pediatric Modification of the Montreal Classification for Inflammatory Bowel Disease: The Paris Classification, Inflamm Bowel Dis, 2011, 1314-1321.

[21] Jan Lotvall, Cezmi A. Akdis, Leonard B. Bacharier, Leif Bjermer, Thomas B. Casale, e Adnan Custovic, f Robert F. Lemanske, Andrew J. Wardlaw, Sally E. Wenzel and Paul A. Greenberger, Asthma endotypes: A new approach to classification of disease entities within the asthma syndrome, American Academy of Allergy, Asthma & Immunology, 2011, 355-360.

[22] S. Husby, S. Koletzko, I.R. Korponay-Szabo, M.L. Mearin, A. Phillips, R. Shamir, A. Ventura and K.P. Zimmer, European Society for Pediatric Gastroenterology, Hepatology, and Nutrition Guidelines for the Diagnosis of Coeliac Disease, European Society for Pediatric Gastroenterology, 2012, 136-160.

[23] K DOI, Current status and future potential of computer-aided diagnosis in medical imaging, The British Institute of Radiology, 2015, 1-17.

[24] Filippo Amato, Alberto López, Eladia María Peña-Méndez, Petr Vaňhara, AlešHampl and Josef Havel, Artificial neural networks in medical diagnosis, J Appl Biomed, 2013, 47-58.

[25] Muthu Rama Krishnan Mookiah, U. Rajendra Acharya, Choo Min Lim, Andrea Petznick and Jasjit S. Suri, Data mining technique for automated diagnosis of glaucoma using higher order spectra and wavelet energy features, Knowledge-Based Systems, 2012, 73–82.
[26] NahlaBarakat, Andrew P. Bradley and M. Nabil Barakat "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus", IEEE, 2010, 1-21.
[27] JuanyingXie and Chunxia Wang, Using Support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases, Expert Systems with Applications, 2010, 1-23.
[28] OrhanEr, FeyzullahTemurtas and A. ÇetinTanrıkulu, Tuberculosis Disease Diagnosis Using Artificial Neural Networks, Springer, 2010, 1-5.
[29] Brant W. Chee, Richard Berlin and Bruce Schatz, Predicting Adverse Drug Events from Personal Health Messages, Adverse drug events, 2011, 1-10.
[30] Shu-Hsien Liao, Pei-Hui Chu and Pei-Yuan Hsiao, Data mining techniques and applications – A decade review from 2000 to 2011, Expert Systems with Applications, 2012, 11303–11311.
[31] Xuezhong Zhou, Shibo Chen, Baoyan Liu, Runsun Zhang, Yinghui Wang, Ping Li, Yufeng Guo, Hua Zhang, Zhuye Gao and Xiufeng Yan, Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support, Artificial Intelligence in Medicine, 2010, 139–152.
[32] Darcy A. Davis, Nitesh V. Chawla, Nicholas A. Christakis and Albert-László Barabási, Time to CARE: a collaborative engine for practical disease prediction, Data Min Knowl Disc, 2010, 388-415.
[33] JavadSalimiSartakhti, Mohammad Hossein Zangooei and KouroshMozafari, Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA), Elsevier, 2011, 1-11.
[34] AbdelghaniBellaachia and Erhan Guven, Predicting Breast Cancer Survivability Using Data Mining Techniques, IEEE, 2011, 1-4.
[35] DivyaTomar and Sonali Agarwal, A survey on Data Mining approaches for Healthcare, International Journal of Bio-Science and Bio-Technology, 2013, 241-266.
[36] K. Sudhakar and Dr. M. Manimekalai, Study of Heart Disease Prediction using Data Mining, International Journal of Advanced Research in Computer Science and Software Engineering, 2014, 1157-1160.
[37] JesminNahar, Tasadduq Imam, Kevin S. Tickle and Yi-Ping Phoebe Chen, Computational intelligence for heart disease diagnosis: A medical knowledge driven approach, Expert Systems with Applications, 2013, 96–104.
[38] Boris Milovic and Milan Milovic, Prediction and Decision Making in Health Care using Data Mining, IJPHS, 2012, 69-78.
[39] Resul Das, A comparison of multiple classification methods for diagnosis of Parkinson disease, Expert Systems with Applications, 2010, 1568–1572.
[40] Jinn-Yi Yeh, Tai-Hsi Wu and Chuan-Wei Tsao, Using data mining techniques to predict hospitalization of hemodialysis patients, Decision Support Systems, 2011, 439–448.