Chronic Disease Detection Model Using Machine Learning Techniques

Vishal Dineshkumar Soni

Abstract: Now-a-days, people face various diseases due to the environmental condition and their living habits. So, the prediction of disease at earlier stage becomes important task. But the accurate prediction on the basis of symptoms becomes too difficult for doctor. The correct prediction of disease is the most challenging task. To overcome this problem data mining plays an important role to predict the disease. Medical science has large amount of data growth per year. Due to increase amount of data growth in medical and healthcare field the accurate analysis on medical data which has been benefits from early patient care. With the help of disease data, data mining finds hidden pattern information in the huge amount of medical data. Data mining is an essential phase in exploring information in libraries where intelligent tools are used to identify trends. Data mining is an important phase in exploring information in libraries where clever tools are used to identify trends. Breast cancer risk has been shown in India to develop 1 in 28 women using the precise classification to test the breast cancer data with a total of 569 rows and 32 columns. Similararly we use Heart disease dataset and Lung cancer dataset , In order to build reliable prediction models for these chronic dieseases using data mining techniques, we are evaluating data accessible from the UCI deep learning data collection in Wisconsin. In this experiment, we compare four results of disease classification techniques with genetic clustering and comparison of the results show that sequential minimal optimization (SMO) has a higher accuracy, i.e. 99.61 percent.

Keywords: SVM, Machine learning, Disease prediction, Medicine, and Accuracy

I. INTRODUCTION

Ever since the first computer or IT technology was introduced, there has been exponential growth with innumerable applications being constantly explored. [1]. Artificial intelligence is being used in multiple areas like cyber security [2] and Disease prediction. We study multiple disease model with chronic disease like Lung cancer, Breast cancer and Diabetes. A typical benign mass has a circular boundary, which is circumscribed and round, but a malignant tumour usually has a raw and blurry boundary that is speculated. [3]. There are numerous kinds of Machine Learning Techniques like Unsupervised, Semi Supervised, Supervised, Reinforcement, Evolutionary Learning and Deep Learning. In many areas and many applications, the resolution of such a problem is based on the processing of features extracted from the original images acquired in the real world and structured as vectors. The quality of the processing system depends highly on the right choice of the constitution of these vectors. But in many cases, the resolution of the problem becomes almost impossible because of the excessive dimensionality of these vectors or inconsistencies which can appear in the data. Therefore, it is often useful and sometimes necessary to reduce the dimension of the dataset samples to a more compatible dimension, even if this reduction may lead to a small loss of information. Accurate and precise evidence will be collected to support the doctor's detection and treatment of disease, both healthy and malignant, at an early stage for an exact model. This would save the doctors time and boost performance. This paper primarily deals with how the diagnosis of various disease using machine learning. Many tools are required for the analysis and detection of patterns in breast cancer. Data set is partitioned into two or more than two classes. Such classifiers are utilized for medical data investigation and disease prediction. Traditional SVM training typically requires a QP package and this package takes more time to solve the QP optimization problem, in particular for large dataset problems. SVM preparation is thus sluggish, and the preparation of broad data sets requires time. The SMO-SVM approach minimizes memory storage, is more precise, and quicker to implement [4]. We propose a hybrid SMO-GC (Sequential minimal optimization with genetic clustering) method to produce an optimum result on Disease Dataset of Breast cancer, Lung cancer and Diabetes

II. LITERATURE REVIEW

Machine learning methods for the theoretical study of life are commonly applicable. Numerous researches based on the value of medical diagnostic technologies have been published. Such experiments have applied various solutions to the issue and obtained a precise classification precision [5] Using an artificial cortical network to determine treatment in breast cancer. They also checked their method on a small data collection, but the findings indicate that they grasp the real survival. And al. [6] Even, in the case of patients with breast cancer, the usage has been made of a naïve bay, decision tree and neural backpropagation network. While the findings obtained were strong (about 90 per cent accuracy), they were not necessary because the data were divided into two groups: one for more than five years of survival and the other for those who died within five years. The findings were not relevant. [7] An approach to pick applications focused on the usability evaluation of function selectors. This strives to have a simple, consistent set of functions without missing the predictive precision dimension. A rating algorithm is used to assign confidence to characteristics. [8] By using fuzzy reasoning to reduce the scale of the initial problem, the suggested hybrid GA / SVM solution. A sub-set of healthy genes is to be identified, which are then tested by SVM. [9] This study aimed at comparing the performance of the Artificial Neural Network (ANN) and Vector Machine Support (SVM) for the classification of liver cancer. In terms of accuracy, flexibility, specificity and the efficiency of both models were contrasted and validated on BUPA Liver Disorder Dataset. Area under Curve (AUC). [10] Used in combination with the prediction of heart disease in association mining and genetic algorithm. The suggested methodology used the genetic algorithm mutation Gini index statistics for the interaction method and crossover. They have employed an increased quality function collection methodology. [17] Designed the Alzheimer disease risk prediction system with the help of EHR data of the patient. Here they utilized active learning context to solve a real problem suffered by the patient. In this the active patient risk model was build. For that active risk prediction algorithm is utilized the risk of Alzheimer disease. N. H. Barakat, [18] proposed an intelligible support vector machines for diagnosis of diabetes mellitus, etc. The classifiers should be used for diabetes prediction and they are recommended to improve them through the production of hybrid models

III. CLASSIFICATION

Classification is one of the strategies for data mining explicitly used to evaluate a specific dataset, takes every instance and assigns it to a specific class [11]. This procedure is structured to eliminate classification mistakes. The classification allows extraction of models that define the classes for a specific set of data. In knowledge processing, three separate research methods are used: directed and unregulated research. [12]. Classification is a preliminary step in the process of analyzing several cases. The final reason is to improve domain understanding or predictions. Through research, there have been suggested several various forms of classification strategies. One of the core activities of data mining is the development of reliable classification systems. For data mining and machine learning science, this is a critical mission. Throughout literature that incorporates decision-making trees, naive- Bayesian processes, serial minB.B.F.um optimization (SMO), IBK, B.F. tree, etc., several various forms of classification strategies were mentioned.

IV. FEATURE SELECTION

The increased use of computers from all points of view leads to the collection of large data. These data are vast and systematically connected data in order to define the appropriate trends, such that data mining is a critical field for data analysis, estimation and other activities. It has entered a dynamic field in research to resolve the theoretical issues in real time. In several places where data processing operations are required, big data mining is used. These data mining and development strategies were widely used at different level such as pattern recognition etc., and the collection of apps in virtually every area plays an important role. The purpose of the selection is to determine the smallest subset of characteristics possible. The framework selects the base of original features by deleting obsolete and unnecessary data dimensionality features without missing any usable details. Before data mine tasks are introduced, it is the crucial pre-processing phase. The precision of mines, the time of calculation and the understanding of the test are enhanced. Three types of selection methods consist of filters, wrappers and embedded approaches.

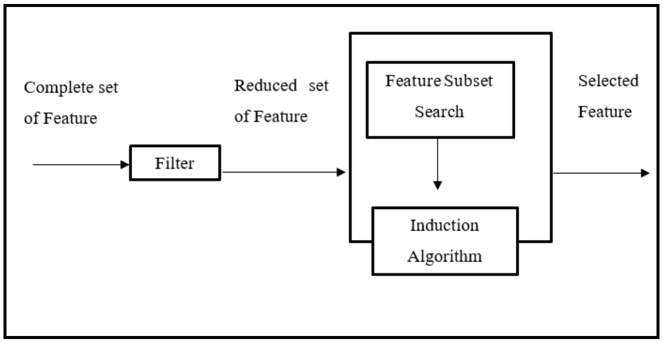


Figure 1: *Hybrid feature selection method [16]*

As Discussed in [16] the Filter selects the function without depending on the type of classifier used. The advantage of this method is that the selection of features must be done only once, it is simple and regardless of the kind of classifier used, but this procedure lacks the relationship with the classifier, every feature is interpreted separately from functional dependence. The approach of Wrapper depends on the classification used. The classifier results are used to assess the goodness of the feature or attribute specified. Another approach has the benefit that the filtering cycle eliminates the downside which is simpler than the filtering system as it still takes all the dependencies. The next embedded approach is to combine a filter algorithm with a wrapper approach in order to find an optimal feature sub-set integrated in the classification structure. The advantage of this method is less costly and less susceptible than the wrapper approach. During the last two decades, several Various applications of feature selection:

- Text mining
- Image processing and computer vision
- Industrial applications
- Bioinformatics

V. MATERIAL AND METHODS

Weka is a platform for data mining that uses a series of algorithms. Such algorithms may be explicitly added to the data or from the Java code-named.

“Weka is a resource set for:

- Regression.
- Clusters.
- Association.
- Pre-processing data.
- Classification Rating.
- Visualization.”[13]

We have classifiers in Weka to estimate quantitative or numerical quantities. Decision and lists are available for learning systems, vector support computers, classifiers dependent on cases, logistic regression and Bayes' networks. All tabs are enabled once the data is loaded. According to the specifications, checks and mistakes, the right algorithm for simple data representation can be sought. The Cluster tab helps the individual to classify clusters or classes of incidents in the data collection. Clustering will provide the consumer with data for review. The training set, percentage division, test set and classes given are used for clustering, for which users may disregard other attributes of the data set on the basis of requirements. “K-Means, EM, Cobweb, X-means and FarthestFirst are included in Weka.”

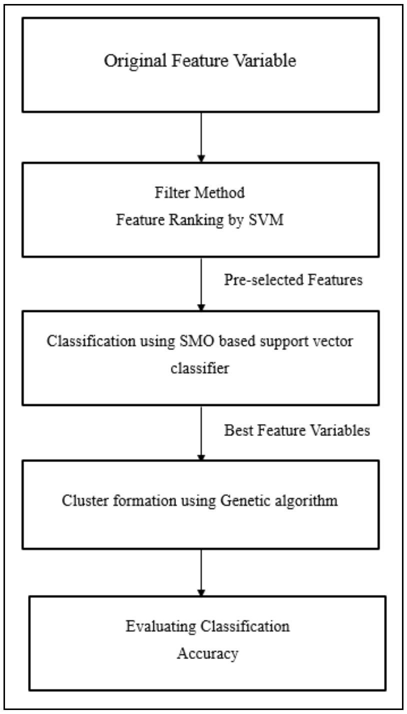
A. PROPOSED METHOD

1) Sequential Minimal Optimization (SMO)

The new algorithm for training supporting vector machines (SVMs) is sequential minimum optimization (SMO). John Platt is the algorithm of minimum sequential optimization (SMO) in 1998 [14] was a quick and simple method for SVM training. The underlying theory is for the dual quadratic optimization to be solved by optimizing the total sub-set of two components. The benefit of SMO is that the presentation is quick and logical. Introducing. A broad issue of quadratic programmatization is the learning of a vector supporting computer. This central question of quadratic programming is separated by SMO in a variety of small potentials. Such a limited square problem system, which prevents the usage of time-consuming quadratic numerical programming as an internal loop, has been analytically resolved. SMO has a linear memory function that allows SMO to perform extensive workouts. Since it is avoided to measure the matrix, the regular SMO scales for various research problems like linear and quadratic chunking SVM algorithm scales like linear and cubic. SMO time is controlled by SVM assessment; SMO is also the fastest for linear and sparse data sets.

B. FRAMEWORK

An illustrative description of SVM and the clustering model dependent on genes is illustrated in Fig.1. Below is an algorithm for the improved SVM method for genetic clustering.


Original Feature Variable
Filter Method
Feature Ranking by SVM
Pre-selected Features
Classification using SMO based support vector classifier
Best Feature Variables
Cluster formation using Genetic algorithm
Evaluating Classification Accuracy

Proposed Framework

Initial variables which will be filtered by SVM filter are chosen. The performance is based on sensitivity, specificity, measurement, reminder, accuracy. In the data mining industry, these metric measures played a crucial role in evaluating the outcomes of various classifications and used to guideline the algorithms, as seen in Table 3.2.

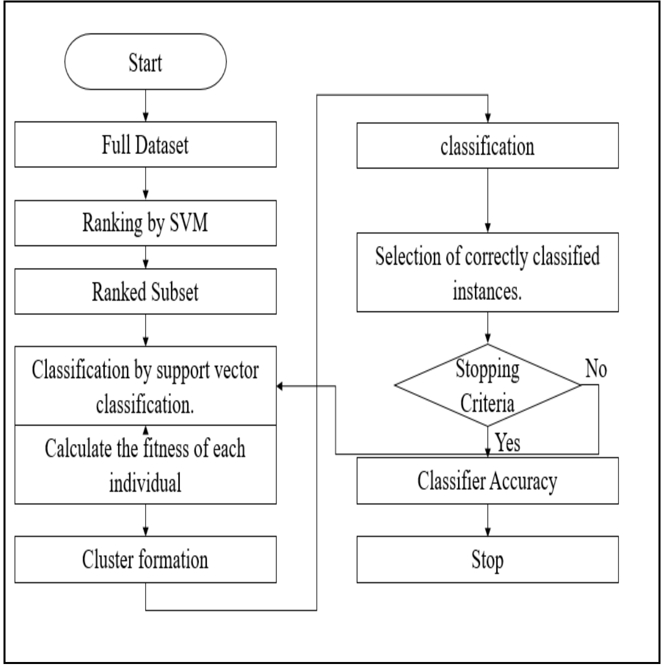
TP = true positives: number of examples predicted positive that are actually positive
FP = false positives: number of examples predicted positive that are actually negative
TN = true negatives: number of examples predicted negative that are actually negative
F.N. = falsifying negatives: the following metrics are measured in terms of the number of instances of negative predictions which are good classifier results. The classification efficiency is measured in the following indexes.

Evaluation parameters

S.No	Metrics	Formula	Evaluation focus
1	Accuracy	(TP+TN) / (TP+TN+FP+FN)	Measures the ratio of correct predictions over the total number of instances evaluated
4	Precision	TP / (TP+FP)	Measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class.
5	Recall	TP / (TP+TN)	Measure the fraction of positive patterns
6	F-Measure	F2 * (precision * recall) / (precision + recall)	Represents the harmonic mean between recall and precision values

C. The proposed hybrid feature selection algorithm

Within this portion, we define a basic genetic clustering SMO algorithm as suggested by our experiments. Several operations have to be calculated before we can use the genetic algorithm. It is the first population, fitness, selection and crossover.


Start
Full Dataset
Ranking by SVM
Ranked Subset
Classification by support vector classification.
Calculate the fitness of each individual
Cluster formation
classification
Selection of correctly classified instances.
Stopping Criteria
No
Yes
Classifier Accuracy
Stop

Proposed algorithm methodology

STEP 1: Population initialization: The initial population is randomly generated. A random function (g) produces a random floating number in [0 , 1] for each piece. If the number exceeds a threshold value, g = 1. If not, g = 0. For eg, the value of the threshold may be set to 0.5 to equate 1 or 0.

STEP 2: Sequential minimum design for vector recognition help training.

STEP 3: This specification removes all incomplete values internationally and transforms them into binary attributes. These features are often uniform by nature. (This is critical in the classification analysis of performance coefficients based on normalized data, not the original data.

STEP 4: The genetic algorithm incorporates K-means with a genetic algorithm into a modern statistical clustering system.

VI. RESULTS AND DISCUSSION

In our opinion, genetic algorithms are original methods which can be used for extraction of functions. They suggest a system for the evaluation of individuals based on the combination of SVM[15] classifiers that have been qualified for each function. More specifically, we associate an SVM classifier with each primitive and implement a selection of classifiers. In the method we propose, classifier learning is carried out in one step before the genetic algorithm for each primitive. We depend on a mixture of these groups to determine the genetic algorithm fitness function. Therefore, we significantly reduce the training period for the rising classifier. The suggested selection method, therefore, decreases the implementation period dramatically.

A. SMO based SVM algorithm results

We ran SMO classifier on filter output with 5K fold. Various other algorithms were compared with our method and given outputs can be seen below.

Table 2*:Result on accuracy with correctly and incorrectly classified instances.*

	ACCURACY	CORRECTLY CLASSIFIED INSTANCES	INCORRECLTY CLASSIFIED INSTANCES
Lung cancer	81.8182	108	24
Breast cancer	99.8243	568	1
Diabetes	78.9272	206	55

	LUNG CANCER	Breast cancer	Diabetes
Precision	1.000	1.000	0.812
Recall	0.333	0.995	0.899
F-Measure	0.500	0.998	0.853
ROC Area	0.667	0.998	0.727

Where B=Benign and M=Malignant.

Confusion Matrix produced for SMO classifier

a	b	
201	11	a=M
2	355	b=B

B. Genetic clustering algorithm results

It was developed by [15]. Standard ManhattanDistanceclass is used for similarity measure in the results. Use of fundamental missing values is done through SimpleKMeans. If a process yields a chromosome with all documents belonging to one cluster, chromosomes are modified before at least two clusters are identified. The selection process is used to copy a chromosome into the new composition based on the values of the function to optimize it. This implies offering more likelihood of passing to the next generation to chromosomes whose health feature value is higher.

Time taken to build model (full training data)	38.81 seconds.

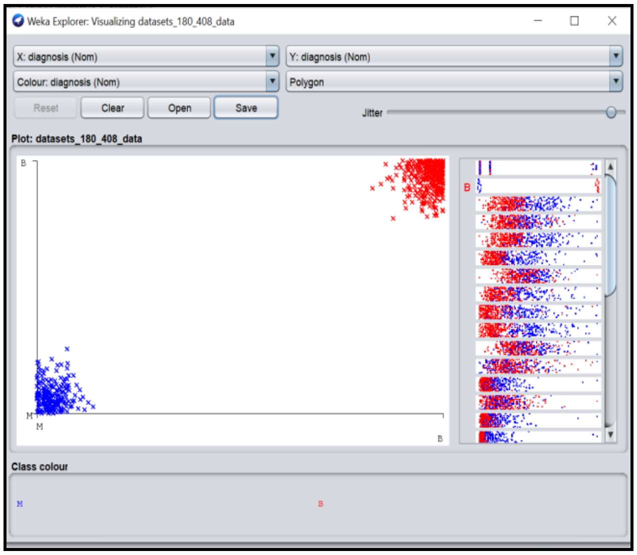
Here it can be seen that the algorithm took 38.81 seconds to execute.

Number of instances based on cluster formation

Confusion matrix of cluster.

0	1	
184	28	M
8	349	B

In the end it was concluded that total incorrectly clustered instances were 36 which were 6.32% of the total data set.Visualization on cluster formation can be seen in the given figure below which shows the assignment of cluster formation.



Weka Explorer: Visualizing datasets_180_408_data

X: diagnosis (Nom)

Y: diagnosis (Nom)

Colour: diagnosis (Nom)

Polygon

Plot: datasets_180_408_data

Blue color in the graph shows Malignant instances and Red color shows the Benign cases in the Dataset.

VII. CONCLUSION

A modern approach to function selection is introduced in this article. This method of selection is based on genetic algorithms combined with the SVM classification. Our method has been tested with Various other classifier of different types on the datasets for breast cancer, Lung cancer and Diabetes. The results demonstrate the robustness of the approach proposed.

VIII. REFERENCES

[1] Soni, Vishal Dineshkumar, Information Technologies: Shaping the World under the Pandemic COVID-19 (June 23, 2020). Journal of Engineering Sciences, Vol 11, Issue 6,June/2020, ISSN NO:0377-9254; DOI:10.15433.JES.2020.V11I06.43P.112 , = Available at SSRN: https://ssrn.com/abstract=3634361

[2] Soni, Vishal Dineshkumar, Challenges and Solution for Artificial Intelligence in Cybersecurity of the USA (June 10, 2020). Available at SSRN: https://ssrn.com/abstract=3624487 or http://dx.doi.org/10.2139/ssrn.3624487

[3] American college of radiology, Reston VA, Illustrated Breast imaging Reporting and Data system (BI-RADSTM) , third edition, 1998.

[4] Urmaliya, Ajay & Jyoti, Singhai. (2013). Sequential minimal optimization for support vector machine with feature selection in breast cancer diagnosis. 2013 IEEE 2nd International Conference on Image Information Processing, IEEE ICIIP 2013. 481-486. 10.1109/ICIIP.2013.6707638.

[5] Bittern, R., Dolgobrodov, D., Marshall, R., Moore, P., Steele, R., AND Cuschieri, A. artificial neural networks in cancer management. e-Science All Hands Meeting 19 (2007), pp.251 – 263.

[6] Bellaachia, A., AND Guven, E. Predicting breast cancer survivability using data mining techniques.

[7] Afef Ben Brahim, Mohamed Limam, 2017, "Ensemble feature selection for high dimensional data: a new method and a comparative study", International Federation of Classification Societies (IFCS), vol. 12(4), pp 937-952

[8] Huerta, E.B., 2006, “A Hybrid GA/SVM approach for gene selection and classification of microarray data”, springer, vol.3907 pp. 34–44

[9] Sallehuddin, R.N.H.aidillah, S.H., Mustaffa, N.H., 2014, “Classification of liver cancer using artificial neural network and support vector machine” In: Proceedings of International Conference on Advance in Communication Network, and Computing, Elsevier Science, pp 487-493.

[10] Jabbar, M.A., Deekshatulu, B.L., Chandra, 2012, “Heart disease prediction system using associative classification and Genetic Algorithm”.

[11] S. S. Nikam, “A Comparative Study of Classification Techniques in Data Mining Algorithms”, Oriental journal of computer science & Technology, vol.8, pp.13-19, April 2015.

[12] S. Neelamegam, E.Ramaraj, “Classification algorithm in Data mining: An Overview”, International Journal of P2P Network Trends and Technology (IJPTT), vol.4, pp.369-374, September 2013.

[13] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[14] Platt, J.C.: Sequential minimal optimization: a fast algorithm for training support vector machines. Technical Report MSR-TR-98- 14, Microsoft Research, 1998.

[15] Islam, M. Z., Estivill-Castro, V., Rahman, M. A., Bossomaier, T. (2018). Combining K-Means and a Genetic Algorithm through a Novel Arrangement of Genetic Operators for High Quality Clustering. Expert Systems with Applications. 91:402-417.

[16] Sangaiah, Ilangovan & Vincent Antony Kumar, A.. (2019). Improving medical diagnosis performance using hybrid feature selection via relieff and entropy based genetic search (RF-EGA) approach: application to breast cancer prediction. Cluster Computing. 22. 10.1007/s10586-018-1702-5.

[17] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, “A relative similarity based method for interactive patient risk prediction,” Springer Data Mining Knowl. Discovery, vol. 29, no. 4, pp. 1070–1093, 2015.

[18] N.H. Barakat, A.P. Bradley, M.N.H. Barakat. (2010). Intelligible SupportVector Machines for Diagnosis of Diabetes Mellitus. IEEE Transactions on Information Technology in Biomedicine. 14(4), pp.1114-1120.