INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume3, Issue2)

Cad Diagnosis Using PSO, BAT, MLR and SVM

Mettildha Mary .I
Department of Information Technology, Sri Ramakrishna Engineering College, Coimbatore, Tamil Nadu
mettilda.lawrance@srec.ac.in

Ilakkiya .M
Department of Information Technology, Sri Ramakrishna Engineering College, Coimbatore, Tamil Nadu
ilakkiya.1305030@srec.ac.in

Kavya .S
Department of Information Technology, Sri Ramakrishna Engineering College, Coimbatore, Tamil Nadu
kavya.1305045@srec.ac.in

Hinduja .R
Department of Information Technology, Sri Ramakrishna Engineering College, Coimbatore, Tamil Nadu
hindhuja.1305029@srec.ac.in

Abstract: Coronary artery disease (CAD) is a most common type of heart disease. CAD happen when blood clot cuts off the heart's blood supply, causing permanent heart damage. Diagnosis of CAD can be done using angiography, echocardiogram, electrocardiogram, which are complex methods. Therefore, studies are done to predict CAD using machine learning algorithms. This study proposes, feature selection by particle swarm optimization(PSO) and Bat algorithms, clustering using K-means and classification using Multinomial logistic regression (MLR) and support vector machine (SVM) algorithms. This technique is cross checked upon 14 attributes with 303 instances. A benchmark dataset from Cleveland heart disease data is used. The Bat-SVM model achieves highest prediction accuracy of 97 %. The proposed model has an increased accuracy from the existing systems.

Keywords: ***Classification, Clustering, Diagnosis, Feature reduction, Heart disease, Machine learning, Optimization.***

1 INTRODUCTION

Coronary Artery Disease happens when the arteries that supply blood to heart muscle become hardened and narrowed. This is due to the build-up of cholesterol and other materials called plaque, on their inner walls. This build-up material is called atherosclerosis. As it grows, less blood can flow through the arteries. This can lead to chest pain (angia) or a heart attack. Most heart attack happens due to this problem.

Over time, CAD can also weaken the heart muscle and contribute to heart failure and arrhythmias. Heart failure means the heart cannot pump blood well to the rest of the body. Arrhythmias are changes in the normal beating rhythm of the heart.

For the past two decades, it has been possible to estimate CAD risks by using of regression equations derived from observational studies and the present study demonstrates similar results, predicting later CAD in middles aged population sample. Prediction models have typically been based on the logistic function although the Weibull distribution has been used. Formulation have often included age, sex, blood pressure, TC, HDLC-C, smoking, diabetes and left ventricular hypertrophy. The prediction od CAD has taken the form of gender specific equation.

The prediction algorithms have been adapted to simplified forms to estimate the CAD risk in patients. The sample used for this report consists of 14 features and 303 instances. The family history for heart disease, physical activity and obesity are not included because these factors work to a large extent through the major risk factor and their unique contribution to CAD prediction can be difficult to quantify.

2 METHODOLOGY

In the past years, medical data mining has become a popular data mining subject. This methodology involves three major data mining techniques: Feature reduction, Clustering and classification.

Machine learning methodology is associate with classification. Machine Learning techniques can extract flexible and comprehensive knowledge from large data set. They also require knowledge for their effective use, but are less complicated to employ and their results are more comprehensible to users.

Many data mining methods involve with mathematical programming techniques. Optimization can contribute to data mining in one of the two ways: (1) Optimization can be component of a larger DM process (2) New DM techniques

Can be built using entirely optimization-based method, which is also called as Optimization Based Approach (OBA) data mining.

Cluster analysis or clustering is the task of grouping a set of objects in such a way that the objects in the same group (called a cluster) are more similar to each other than to those in other groups. It is a main task of explanatory data mining and a common technique for statistical data analysis, used in many fields, including pattern recognition, bioinformatics, and information retrieval. Machine learning techniques deals with the mix of quantitative, qualitative, missing or noisy data so common on engineering. There are various algorithms available for optimization and classification. Machine learning algorithms are described as either supervised or unsupervised. The distinction is drawn from hoe the user classifies the data. In supervised learning, which is referred also to classification, the classes are labelled. These classes can be conceived of as a finite set which is previously arrived.

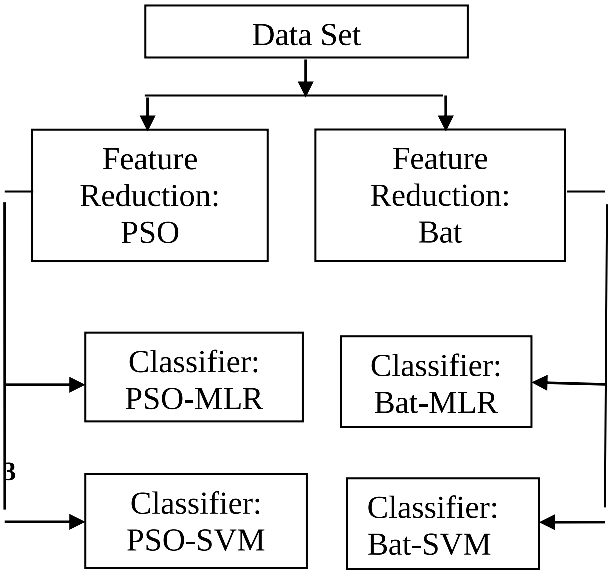

Fig.1.1 Flow Diagram of the proposed model

3DATA SET INFORMATION

The classification task in this database is to determine the presence of heart disease in the patient. It is a integer, valued from 0 to 2. The database contains 14 attributes. All attributes are numeric valued.

Attribute Information:

1. Age: Age in years
2. Sex: 1=male; 0=female
3. Cp: Chest pain type
    - Value1: Typical angina
    - Value2: atypical angina
    - Value3: Non-angina pain
    - Value4: Asymptomatic
4. Tresbps: Resting blood pressure (in mm Hg)
5. chol: serum cholesterol in mg/dl
6. fbs: Fasting blood sugar
    - True=1
    - False=0
7. Restecg: Resting electrocardiogram result
    - Value 0: normal
    - Value 1: have ST-T wave abnormality
    - Value 2: showing probable or definite left ventricular hypertrophy
8. thalch: maximum heart rate achieved
9. exang: exercise induced angina
    - Value 1=yes
    - Value 0=no
10. oldpeak: ST depression induced by exercise relative to rest
11. Slope: slope of the peak exercise ST segment
    - Value 1: upsloping
    - Value 2: flat
    - Value 3: down slopping
12. ca: number of major vessels
13. thal: thalassemia
14. num: diagnosed value

4 OPTIMIZATION ALGORITHMS

A. Particle Swarm Optimization

The particle swarm optimization (PSO) is a computational method that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. It solves a problem by having a population of candidate solutions, here dubbed particles, and moving these particles around in the search-space according to simple mathematical formulae the particle's position and velocity.

Each particle's movement is influenced by its local best known position, but is also guided toward the best known positions in the search-space, which are updated as better positions are found by other particles. This is expected to move toward the best solutions.

The underlying phenomenon of PSO is that knowledge is optimized by social interaction in the population where thinking is not only personal but also social. PSO is based on the principle that each solution can be represented by a vector xi= (xi1, xi2… xiD), where D is the dimensionality of the search for the optimal solutions. Therefore, each particle has velocity, which is represented bas vi= (vi1, vi2… viD). During the movement, each particle updates its position and velocity according to its own experience and that of its neighbours. The best previous position of the particle is recorded as the personal best pbest, and the best position obtained by the population so far is called the gbest. Based on the pbest and gbest values, PSO searches for the optimal solutions by updating the velocity and the position of each particle according to the velocity and the position of each particle according to the following equations:

xt+1 = xt + vt+1 – (1)
vt+1 = (w * vt) + (c1 * r1*pid – xt)+ (c2 * r2 * pgd – xt)

Where t represents the t th iteration in the evolutionary process. d € D represents the dth dimension in the search space. W is the inertia weight, which is to control the impact of the
Previous velocities and current velocities.c1 and c2 are the acceleration constants. R1 and r2 are the random values uniformly distributed in [0, 1] pid and pgd denotes the elements of pbest and gbest in the dth dimension. The velocity is limited by a predefined maximum velocity, vmax and vt+1 € [-vmax, vmax]. The algorithm stops when a predefined criterion is met, which could be a good fitness value or a predefined maximum number iterations.

The choice of PSO parameters can have a large imoact on optimization performance. Selecting PSO parameters that yield good performance has therefore been the subject of much research. The PSO parameters can also be tuned by using another overlaying optimizer, a concept known as Meta optimization. Parameters have also been tuned for various optimization scenarios.

B. Bat Algorithm

The Bat algorithm is metaheuristic algorithm for global optimization. It was inspired by the echolocation behaviour of micro bats, with varying pulse rates of emission and loudness. The idealization of the echolocation of micro bats can be summarized as follows: each virtual bats flies randomly with the velocity at position(solution) with varying frequency or wavelength and loudness .As it searches and finds its prey, it changes frequency, loudness and pulse emission rate. Search is intensified by a local random walk. Selection of the best continues until certain stop criteria are met. This essentially uses a frequency-tuning technique to control the dynamic behaviour of a swarm of bats, and the balance between exploration and exploitation can be controlled by tuning algorithm-dependent parameters in Bat algorithm.

A detail introduction of metaheuristic algorithm including the Bat algorithm is given by the Yang where a demo program in Matlab/octave is available, while a comprehensive review is carried out by parpinelli and lopes a further improvement is a development of an evolving Bat algorithm (EBA)
With better efficiency.

Each bat is associated with a velocity v t i and a location x t i, at iteration t in a d-dimensional search or solution space. Among all the bats there exists a current best solution x*.therefore, the above three rules can be translated into the updating equations for x t i and velocities v t i.

Fi = fmin + (fmax - fmin) β, (1)
V t I = v t – 1 I + (x t – 1 i – x*) fi, (2)
X t i = x t – 1 I + v t i, (3)

Where β € [0, 1] is a random vector drawn from a uniform distribution. As mentioned earlier, we can either use wavelengths or frequency for implementation, we will use fmin =0 and fmax = O (1), depending on the domain size of the problem of interest. Initially each bat is randomly assigned frequency which is drawn uniformly from [fmin, fmax]. For this reason, bat algorithm can be considered as frequency-tuning algorithm. To provide a balanced combination of exploration and exploitation. The loudness and pulse emission rate essentially provide a mechanism for automatic control and auto zooming into the region with promising solution.

A further key concept in the field of swarm intelligence is stigmergy. Stigmergy is a mechanism of indirect coordination between agents or action. The principle is that the trace left in the environment by an action stimulation is the performance of the next action, by same or different agent. In that way, subsequent action tend to reinforce and build on each other, leading to spontaneous emergence of coherent, apparently systematic activity. Stigmergy is the form of self-organisation. It produces complex, seemingly intelligent structures, without need for any planning, control or even direct communication between the agents. As such it supports efficient collaboration between extremely simple agents, who lack any memory or intelligence or even awareness of each other's.

V.CLASSIFICATION ALGORITHMS

A. Multinomial Logistic Regression

Multinomial Logistic Regression is a classification method that generalizes logistic regression to multiclass problems, i.e. with more than two possible discrete outcomes. That is, it is a model that is used to predict the probabilities of the different possible outcomes of categorically distributed dependent variable, given a set of independent variables (which may real-valued, binary-valued, and categorical-valued).

The statistical classification problem have in common a dependent variable to be predicted that comes from one of a limited set of items which cannot be meaningfully ordered, as well as a set of independent variables(also known as features, explanatory), which are used to predict the dependent variable. Multinomial Logistic Regression is a particular solution to the classification problem that assumes that a linear combination of the observed features and some problem specific parameters can be used to determine the probability of each particular outcome of the dependent variable. The best value of the parameters for a given problems are usually determine some training data.

The Multinomial Logistic Regression module assumes that data are case specific; that is, each independent variable has a single value for each case. The Multinomial Logistic model also assumes that the dependent variable cannot be perfectly predicted from the independent variable for any case. As with other types of regression, there is no need for the independent variable to be statistically independent from each other (unlike, for example, in a Naïve Bayes Classifier); However, collinearity is assumed to be relatively low, as it becomes difficult to differentiate between the impact of several variables if this is not a case.

If the multinomial Logistic is used to model choices, it is relies on the assumption of the independent of irrelevant alternatives (IAA) which is not always desirable. This assumption states that the art preferring one class over another do not depend on the presence of relevant or irrelevant alternatives. For example, the relative probabilities of car or a bus to work do not change of a bi-cycle is added as an additional possibilities. This allows a choice of k-alternatives to be modelled as a set of k-1 independent binary choices, in which one alternative is chosen as a "pivot" and the other k-1 compared against it, one at a time. The IAA hypothesis is a core hypothesis in rational choice theory: However numerous studies in psychology show that individuals often violate this assumption when making choices.an example of a problem case arises if choices include a car and a blue bus. Suppose the odds ratio between the two is 1: 1.

Now the option of the red bus is introduces, a person may be indifferent between a red and blue bus, at hence may exhibit a car: blue bus: red bus odds ratio of 1: 0.5: 0.5, thus maintaining a 1: 1 ratio of car: any bus while adopting a changed car: blue bus of ratio 1 : 0.5. Here the red option was not in fact irrelevant, because a red bus was the perfect substitute for the blue bus.

If the Multinomial Logistic is used to model choices, it may in some situation impose too much constrain on the relative preference between the different alternatives. This point is especially important to take into account if the analysis aims to predict how choices would change if one alternative was too disappeared (for instance if one political candidate withdraws from three candidate race). Other models like the nested logit are the multinomial probity may be used in such cases as they allow for violation of IAA.

B.SUPPORT VECTOR MACHINE

In machine learning, support vector machine (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis. Given a set of training example, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new example to one category or the other, making it non probabilistic binary linear classifier. An SVM model is a representation of the example as a points in space, mapped so that the examples of the separate categories are divide by a clear gap that is as wide as possible. New examples are then are mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

In addition to a performing linear classification, SVMs can efficiently performed a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high dimensional feature spaces.

When data are not labelled, Supervised learning is not possible, and an unsupervised learning approach is required which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. The clustering algorithms which provides an improvement to the support vector clustering and is often used in industrial applications either when data are not labelled or when only some data are labelled as a pre-processing for a classification pass.

The effectiveness of SVM depends on the selection of kernel, the kernel's parameters, and soft margin parameter. A common choice is a Gaussian kernel, which has a single parameter. The final model, which is used for testing and for classifying new data, is then trained on the whole training set using the selected parameters.

CONCLUSIONS

Through this study, the best algorithm which could be used to predict the coronary artery disease can be witness. Comparison of each optimization algorithms with the classification algorithms will be made. Based on this comparisons accuracy will be calculated. The algorithm with a greater accuracy can be chosen and proposed as the best algorithm for this disease prediction purpose.

REFERENCES

1. Luxmi Verma, Sangeet Srivastava, P. C. Negi, A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data, Springer Science+Business Media, june 2016

2. Acharya, U.R., Faust, O., Sree, V., Swapna, G.,Martis, R.J., Kadri,N.A., and Suri, J.S., Linear and nonlinear analysis of normal and CAD-affected heart rate signals. Comput. Methods Prog. Biomed.113 (1):55–68, 2014.

3.Giri, D., Acharya, U.R., Martis, R.J., Sree, S.V., Lim, T.C., Ahamed .T., and Suri, J.S., Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform. Knowl.-Based Syst. 37:274–282, 2013.

4. Peter, T. J., &Somasundaram, K., An empirical study on prediction of heart disease using classification data mining techniques. In Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on (pp. 514–518). IEEE, 2012.

5. Bouali, H., & Akaichi, J., Comparative Study of Different Classification Techniques: Heart Disease Use Case. In Machine Learning and Applications (ICMLA), 2014 13th International Conference on (pp. 482–486). IEEE.

6. R. O. Bonow, D. L. Mann, D. P. Zipes, and P. Libby, “Braunwald’s Heart Disease: A Textbook of Cardiovascular Medicine”, 9th edition: New York, Saunders, 2012.

7. S. Dua, X. Du, V.S. Sree, T.V.I. Ahmed, Novel classification ofcoronary artery disease using heart rate variability analysis,Journal of Mechanics in Medicine and Biology 12 (4) (2012),1240017-1-19.

8. Z. Zhao, C. Ma, An intelligent system for noninvasive diagnosis of coronary artery disease with EMD-TEO and BPneural network, in: International Workshop on Education Technology and Training and International Workshop on Geoscience and Remote Sensing, vol. 2, 2008, pp. 631–635.

9. Amin, S. U., Agarwal, K., & Beg, R., Genetic neural network based data mining in prediction of heart disease using risk factors. In Information & Communication Technologies (ICT), 2013 I.E. Conference on (pp. 1227–1231). IEEE, 2013.

10. Kumar, R., Negi, P.C., Bhardwaj, R., Kandoria, A., Asotra, S., Ganju, N., and Marwah, R., Clinical and non-invasive predictors of the presence and extent of coronary artery disease. Indian HeartJ. 66:S28, 2014.