# Feature Selection Using Binary Grey Wolf Optimization Algorithm on Chronic Kidney Disease Dataset

J. Thamil Selvi, G. Soundharamanikandan, M. Swathikha and R.S. Latha

**Abstract---** The Chronic Kidney disease is the most important health issues concerning the people as a whole. Chronic diseases lead to morbidity and increase of death rates in India and other low and middle income countries. The chronic diseases account to about 60% of all deaths worldwide. 80% of chronic disease deaths worldwide also occur in low and middle income countries. In India, the number of deaths due to chronic disease found to be 5.21 million in 2008 and seems to be raised to 7.63 million in 2020 approx 66.7% .In Chronic Kidney Disease dataset contain 24 features and achieved Accuracy. In this paper, a binary Grey Wolf Optimization algorithm is used for feature selection and compared with PSO-KDE and GA-KDE where PSO-KDE model is proposed that hybridize the particle swarm optimization (PSO) and kernel density estimation (KDE) based classifier to diagnosis of chronic kidney disease. Classification performance and the number of selected features are the criteria used to design the objective function of PSO-KDE and GA-KDE. The Experimental results prove that the PSO-KDE model has better average performance in diagnosis of kidney disease.

**Keywords---** Chronic Kidney Disease(CKD), Particle Swarm Optimization, Kernel Density Estimation, Binary Grey Wolf Optimization, Feature Selection.

## I. INTRODUCTION

The massive advancement due to the amount of biological data available has raised a gloomy question of being classified, managed effectively and to be transfer the raw data to meaningful information. The emergence of this colossal amount of calls into question the pattern of modern computation. It can be addressed using data mining algorithms. Machine Learning has application in accounts for the latest progress in the field of bioinformatics, computational biology and application of machine learning methods on prominent problems in human biology and behavior. The notions of supervised and unsupervised learning make this process easy and comprehensible. By simplifying abstraction the statistical predictions of a system is obtained which constitutes a model. As people’s day to day life become more and more modernized and extended life span in the society, Chronic Kidney Disease (CKD) also found common and result in degradation in the functionalities of kidney function. Once any person gets CKD, they may suffer from the disease which may decrease their working capability as well as living quality. It is also rapidly results in other chronic diseases like high blood pressure, anemia, weak bones due to poor nutritional health and nerve damage. In the meantime, kidney disease maximises the patient risk of contracting heart and blood oriented diseases. Chronic kidney disease even causes other chronic disease such as diabetes, high blood pressure and other disorders. High risk groups are classified as person with diabetes, hypertension , and hereditary. It is possible to get rid of chronic kidney disease through early detection

*J. Thamil Selvi, Student, Department of CSE, Kongu Engineering College, Erode, Tamil Nadu. E-mail:thamilselvijaganathan@gmail.com*

*G. Soundharamanikandan, Student, Department of CSE, Kongu Engineering College, Erode, Tamil Nadu.*

*M. Swathikha, Student, Department of CSE, Kongu Engineering College, Erode, Tamil Nadu.*

*R.S. Latha, Assistant Professor, Department of CSE, Kongu Engineering College, Erode, Tamil Nadu.*

and proper treatment once the progress of the disease is observed it may greatly leads to kidney failure[1].

Feature selection provides away for identifying the important features and removing irrelevant (redundant) ones from the dataset[2].The feature selection objectives are data dimensionality reduction, improving prediction performance, and good data understanding for different machine learning applications[3].In the real world applications, data representation often uses too many features with redundancy features, which means certain features can take the role of another and the unnecessary features can be removed. Moreover, the relevant (interdependence) features have an influence on the output and contain important information that will be obscure if any of them is excluded [4].

Previously, an exhaustive search for the optimal set of features(attributes) in a high dimensional space may be unpractical. Many researches try to model the feature selection as a combinatorial optimization problem, which the set of features lead to the best feature space separability [5].The objective function can be the classification accuracy or some other criterion that might consider the best trade-off between attribute extraction computational burden and efficiency [6].

Particle swarm optimization (PSO) algorithm is an evolutionary computing technique which is powerful and computationally efficient[7]. Recently, it has been successfully applied in large number of applications and difficult optimization problems [8]. Yeh et al.[8] proposed an approach using statistical method and discrete PSO for mining breast cancer pattern. Wang et al. [9] presented an integration of rough sets and particle swarm intelligence for the selection of high quality feature subsets. Xue et al. proposed a PSO based feature selection approach and showed that PSO is effective search technique for feature selection problem. Ouyang et al. [10] presented a hybrid particle swarm optimization to estimate the Muskingum model parameters. A combination of multi-objective PSO with pare to optimal solutions was presented by Jia et al. [11] to solve the batch processes problem. Ouyang et al. [12] presented a parallel hybrid PSO (PHPSO) algorithm to solve a real case of one-dimensional heat conduction equation. Chen et al. [13] applied an improved cooperative PSO to train the feed forward neural network. The problem of cost optimization of mixed feeds was solved by particle swarm optimization. Escalante et al. [14] proposed an application of PSO to the problem of full model selection (FMS) for classification.

Grey wolf optimization (GWO) is a newly introduced evolutionary algorithm, which proposes that the grey wolves have a successful reproduction more than hunting in the pack. Two grey wolves (male and female) have a higher position and managing the other wolves in the pack [15]. In this paper, binary grey wolf algorithm is used for feature selection. Grey wolf optimizer is one of the latest bio inspired techniques, which simulate the hunting process of a pack of grey wolves in nature.

The remainder of this paper is organized as follows: Brief background knowledge on particle swarm optimization is presented in Section 2. Section 3 describes non-parametric density estimation. Section 4 describe Grey Wolf Optimization algorithm. Results and discussion are presented in Section 5.conclusion described in section 6

## II. PARTICLE SWARM OPTIMIZATION

Particle swarm optimization [16] is a heuristic global optimization method that simulates social behavior of bird flocks to a promising position to achieve precise objectives in a multidimensional space. PSO utilizes a population (called swarm) of particles in the search space [17,18]. The status of each particle is characterized according to its position and velocity. xi= {xi1, xi2, . . ., xid}and the velocity of particle i is represented as *v*i= {*v*i1, *v*i2, . . ., *v*id}.To discover the optimal solution, each particle changes its searching direction according to two factors: The best position of a given particle (pbest) and the best position obtained by the swarm so far (gbest). PSO searches for the optimal solution by updating the velocity and position of each particle according to the following equations [19]:

$$v_{id}^{t+1} = wv_{id}^{t} + c_{1} \ r_1(p_{id} - x_{id}^{t}) + c_2 r_2(p_{gd} - x_{id}^{t}) \quad (1)$$

$$x_{id}^{t+1} = x_{id} + v_{id}^{t+1} \quad (2)$$

where t denotes the iteration in the evolutionary space and d denotes the d th dimension in the search space. w is the inertia weight. $c_1$ and $c_2$ are personal and social learning factors. $r_1$ and $r_2$are random values uniformly distributed within the range [0, 1].$p_{id}$ and $p_{gd}$ represent the pbest and gbest in the dth dimension. The basic process of the PSO algorithm is given as follows:

1. Initialization: Particles are initialized with random positions and velocities.
2. Evaluation: The value of objective function is measured for each particle.
3. Find the pbest: If the value of objective function for particle i is better than the pbest of particle i, the current value of objective function is set as the new pbest of particle i.
4. Find the gbest: If any pbest is better than the gbest, gbest is set to the current value.
5. Update velocity and position: The velocity of each particle is updated according to Eq. (1), and the particle is moved to the next position according to Eq. (2).
6. Stopping criterion: If the number of iterations is met, the algorithm will be stopped; otherwise it will be returned to step 2.

## III. NON PARAMETRIC DENSITY ESTIMATION

Non-parametric methods mainly focus on identifying the past conditions which are similar to the conditions at the prediction time [20]. These methods estimate the density directly from the data without any assumptions about the underlying distribution[21-22].

Let X = {xt}Nt=1 be independent and identically distributed d-dimensional random variables (training data) with an unknown density p(.).ˆp(.) is the estimator of p(.) which counts the percentage of the observations which are close to the point x. Non-parametric density estimation is defined as:

$$p = \frac{1}{h} [\frac{No.\{x^{t} \leq x+h\} - No.\{x^{t} \leq x\}}{N}] \quad (3)$$

where No . {c} is the number of elements for which c is true, N is the number of instances, h is the length of the interval, xt is the training instance and x is a new arrival data. No.{xt≤x + h} − No.{xt≤x}denotes the number of training instances that fall in the same interval as x.

The simplest non-parametric estimator is the histogram where divide the input space into a number of bins [23,24]. The histogram requires two parameters to be defined: bin width h and starting position of the first bin x0. Histogram estimate of the probability density function is defined as:

$$\mathrm{p(x)} = \frac{\mathrm{No.\{x^{t}in\ the\ same\ \ bin\ as\ x\}}}{\mathrm{Nh}} \quad (4)$$

The density estimate depends on the starting position of the bins. Naive estimator is an alternative to the histogram for making density estimate. This estimator frees us from setting a starting position of the bins. The naive estimator is defined as:

$$\mathrm{p(x)} = \frac{\mathrm{No.\{x-\frac{h}{2}<x^{t} \leq x+\frac{h}{2}\}}}{\mathrm{Nh}} \quad (5)$$

An alternative way to represent the naive estimator is:

$$\mathrm{p(x)} = \frac{1}{\mathrm{Nh}} \sum_{\mathrm{t=1}}^{\mathrm{N}} \mathrm{w} \frac{\mathrm{(x-x^{t})}}{\mathrm{h}} \quad (6)$$

where the weight function is defined as:

$$w(u) = \left\{ \begin{matrix} 1 & if\ |u| < 1/2 \\ 0 & otherwise \end{matrix} \right\} \quad (7)$$

The naive estimator suffers from being discontinuous, with jumps at xi±h and a derivative of zero everywhere else. Kernel density estimation generalizes the naive estimator to eliminate the discontinuous nature of the resulting probability density function. Kernel density estimator estimates a probability density function for an unknown distribution by summing kernel functions centered at each observed data point. A well-known non-parametric kernel estimator of the density function is the Parzen windows, defined as:

$$p(x) = \frac{1}{Nh} \sum_{t=1}^{N} K(x - \frac{x^t}{h}) \qquad (8)$$

h is the kernel bandwidth or smoothing constant which controls the degree of smoothing and K(.) is the kernel function. To ensure that p(x) is a density, the kernel function should satisfy k(x) ≥ 0 for all x and −∞+∞k(x) dx = 1.A well-known non-parametric kernel estimator of the density function satisfying all of the above properties is the Gaussian kernel: [25]

$$K(u) = \left(\frac{1}{\sqrt{2\pi}}\right)^d \exp[-\frac{\|u\|^2}{2}] \qquad (9)$$

where u denotes the Euclidean distance

## IV. Binary Grey Wolf Optimization

In this approach BGWO the main updating equation can be formulated as shown in Eq. (13)

$$x_i^{(t+1)} = \text{crossover}(x1, x2, x3) \quad (10)$$

where Crossover (x,y,z) is suitable crossover between solutions x,y,z and x1,x2,x3 are binary vectors representing the effect of wolf move towards the alpha, beta, delta grey wolves in order. x1,x2,x3 are calculated using Eqs. (11),(14),and(17), respectively. [ 26]

$$x_1^d = \begin{cases} 1 & if \ \ (x_\alpha^d + bstep_\alpha^d) \geq 1 \\ 0 & otherwise \end{cases} \qquad (11)$$

where dxα is the position vector of the alpha wolf in the dimension d, and bstepd$_\alpha$ is a binary step in dimension d that can be calculated as in Eq. (12).

$$bstep_\alpha^d = \begin{cases} 1 \ if \ cstep_\alpha^d \geq rand \\ 0 \qquad otherwise \end{cases} \qquad (12)$$

Similarly,

$$x_d^2 = \begin{cases} 1 \ if \ (x_\beta^d + bstep_\beta^d) \geq 1 \\ 0 \qquad otherwise \end{cases} \qquad (13)$$

$$x_3^d = \begin{cases} 1 \ if \ \ (x_\delta^d + bstep_\delta^d) \geq 1 \\ 0 \qquad otherwise \end{cases} \qquad (14)$$

Algorithm1. Binary grey wolf optimization algorithm

```
input: n Number of grey wolves in the pack
NIter Number of iterations for optimization.
output: $x_\alpha$ Optimal grey wolf binary position
f ($x_\alpha$) Best fitness value.
1.Initialize a population of n wolves positions at random €[0,1]
2. Find the α, β, δ solutions based on fitness.
3. while Stopping criteria not met do
For each Wolf$_i$ € pack do
Calculate x1,x2,x3 using equation (11),(13),(14)
$^iX_{t+1}$ ←crossover among x1,x2, x3
end
I Update a ,A, C
II Evaluate the positions of individual wolves
III Update α,β, δ
end
```

## V. Result and Discussion

### *Dataset*

Dataset is a collection of data or a single statistical data where every attribute of data represents variable and each instance has its own description.. The datasets used by us contains 25 attributes and 400 instances out of which 250 are suffering from the disease and 150 are not suffering from the disease

## VI. Conclusion

In this work, PSO-KDE give best accuracy and achieve accuracy with minimum number of features and error rate is also minimized.

In future work, BGWO-KDE algorithm can be applied to variety of application and can be compared with other algorithm

## References

[1] A. Dubey, “A Classification of CKD Cases Using Multivariate K-Means Clustering”, International Journal of Scientific and Research Publications, Vol.5, No.8, Pp.1-5, 2015.

[2] B. Chizi, L. Rokach and O. Maimon, “A survey of feature selection techniques”, Encyclopedia of Data Warehousing and Mining, seconded, IGIGlobal, Pp. 1888–1895, 2009.

[3] G. Chandrashekar and F. Sahin, “A survey on feature selection methods”, Comput. Electr. Eng, Vol.40, No.1, Pp.16–28, 2014.

[4] D. Bell and H. Wang, “A formalism for relevance and its application in feature subset selection”, Mach. Learn, Vol.41, No.2, Pp.175–195, 2000.

[5] R.O. Duda, P.E. Hart and D.G. Stork, Pattern Classification, seconded, Wiley-Inter science Publication, USA, 2000.

[6] R.Y.M. Nakamura, L.A.M. Pereira, K.A. Costa, D. Rodrigues, J.P. Papa and X.S. Yang, “BBA: A binary bat algorithm for feature selection”, Graphics:Patterns and Images (SIBGRAPI), Pp.291–297, 2012.

[7] C.L. Huang and J. Dun, “A distributed PSO–SVM hybrid system with feature selection and parameter optimization”, Appl. Soft Comput, Vol.8, 2008.

[8] W.C. Yeh, W.W. Chang and Y.Y. Chung, “A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method”, Expert Syst. Appl., Vol.36, No.4, Pp.8204–8211, 2009.

[9] X. Wang, J. Yang, X. Teng, W. Xia and R. Jensen, Feature selection based on rough sets and particle swarm optimization, Pattern Recognit. Lett., Vol.28, No.4, Pp. 459–471, 2007.

[10] A. Ouyang, K. Li, T. Truong, A. Sallam and E.H.M. Sha, “Hybrid particle swarm optimization for parameter estimation of Muskingum model”, Neural Comput. Appl., Vol.25, No.7–8, Pp.1785–1799, 2014.

[11] L. Jia, D. Cheng and M.S. Chiu, “Pareto-optimal solutions based multi-objective particle swarm optimization control for batch processes”, Neural Comput. Appl., Vol.21, No.6, Pp.1107–1116, 2012.

[12] A. Ouyang, Z. Tang, X. Zhou, Y. Xu, G. Pan and K. Li, “Parallel hybrid PSO with CUDA for lD heat conduction equation”, Comput. Fluids, Vol.110, Pp.198–210, 2014.

[13] D. Chen, C. Zhao and H. Zhang, “An improved cooperative particle swarm optimization and its application”, Neural Comput. Appl., Vol.20, No.2, Pp.171–182, 2011.

[14] H.J. Escalante, M. Montes and L.E. Sucar, “Particle swarm model selection”, J. Mach.Learn. Res., Vol.10, Pp.405–440, 2009.

[15] S.Shoghian and M. Kouzehgar, “A comparison among wolf pack search and four other optimization algorithms”, World Acad. Sci., Eng. Technol., Vol.6, 2012.

[16] J. Kennedy and R. Eberhart, “Particle swarm optimization”, Proceedings of the IEEE International Conference on Neural Networks, Vol. 4, Pp.1942–1948, 1995

[17] Y. Del Valle, G.K. Venayagamoorthy, S. Mohagheghi, J.C. Hernandez and R.G. Harley, “Particle swarm optimization: basic concepts, variants and applications in power systems”, IEEE Trans. Evolut. Comput., Vol.12, No.2, Pp.171–195, 2008.

[18] D.J. Krusienski and W.K. Jenkins, “Non-parametric density estimation based independent component analysis via particle swarm optimization”, IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings, Vol. 4, 2005, March.

[19] Y. Shi and R. Eberhart, “A modified particle swarm optimizer”, Proceedings, IEEE World Congress on Computational Intelligence in: Evolutionary Computation, Pp. 69–73, 1998.

[20] N.T. Ratrout and U. Gazder, “Factors affecting performance of parametric and non-parametric models for daily traffic forecasting”, Procedia Comput. Sci., Vol.32, Pp.285–292, 2014.

[21] E. Alpaydin, Introduction to Machine Learning, second ed, MIT press, 2010.

[22] T. Brox, B. Rosenhahn, D. Cremers and H.P. Seidel, “Non-parametric density estimation with adaptive, anisotropic kernels for human motion tracking”, Human Motion-Understanding, Modeling, Capture and Animation, Springer, Berlin, Heidelberg, Pp. 152–165, 2007.

[23] A. Elgammal, R. Duraiswami, D. Harwood and L.S. Davis, “Background and fore-ground modeling using non-parametric kernel density estimation for visual surveillance”, Proc. IEEE, Vol.90, No.7, Pp.1151–1163, 2002.

[24] P. Ramachandran and T.J. Perkins, “Adaptive bandwidth kernel density estimation for next-generation sequencing data”, BMC Proceedings, Bio Med Central Ltd., 2013.

[25] R. Sheikhpour, M.A. Sarram and R. Sheikhpour, “Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer”, Applied Soft Computing, Vol.40, Pp.113-131, 2016.

[26] E. Emary, H.M. Zawbaa and A.E. Hassanien, “Binary grey wolf optimization approaches for feature selection”, Neurocomputing, Vol.172, Pp.371-381, 2016.