Abstract: Data mining is extracting the useful information from the database. In this paper, we use the classification method for diabetes diagnosis. Classification is a data mining function that assigns objects in a group to target category or classes. Cat Swarm Optimization (CSO) is one of the new swarm intelligence algorithms for finding the best global solution. Because of density, sometimes the pure CSO takes a long time to congregate and cannot achieve the accurate solution. For solving this problem and improving the convergence accuracy level, we propose a new improved CSO namely ‘Statistical Classifier Cat Swarm Optimization’. First, we add a newb inertia weight to velocity equation and then use an adaptive acceleration coefficient. Second, by using the information of two previous/next dimensions and applying a new feature, we reach to a new position update equation compose the average of position and velocity information. Experimental results for six test functions show that in comparison with the pure CSO, the proposed CSODT can takes a less time to converge and can find the best solution in less iteration.

Keyewords: Data mining, diabetes, Swarm Intelligence, Cat Swarm Optimization, Evolutionary Algorithms

I. INTRODUCTION

Function of Optimization is one of the important fields in the computational intelligence theories. There are many algorithms to find the global and local solutions of the problems. Some of these optimization algorithms were developed based on swarm intelligence.

These algorithms reproduce the creature’s swarm behavior and model into algorithm, such as Ant Colony Optimization (ACO) which imitates the behavior of ants. Particle Swarm Optimization (PSO) which imitates the behavior of birds, Bee Colony Optimization (BCO) which imitates the behavior of bees and the recent finding, Cat Swarm Optimization (CSO) which imitates the behavior of cats.

By simulating the behavior of cats and modeling into two modes, CSO can solve the optimization problems. In the cases of functions optimization, CSO is one of the best algorithms to find the global solution. In comparison with other heuristic algorithms such as PSO and PSO with weighting factor, CSO usually achieves better result. But, because of algorithm complexity, solving the problems and finding the optimal solution may take a long process time and sometimes much iteration is needed.

So in this article, we propose an improved CSO in order to achieve the high convergence accuracy in less iteration. First we use an adaptive inertia weight and adaptive acceleration coefficient. So, the new velocity update equation will be computed in an adaptive formula. Then, our aim is to consider the effect of previous/next steps in order to calculate the current position.

So by using a factor namely ‘Forgetting Factor’, information values of steps will be different. Finally, we use an average form of position update equation composing new velocity and position information.

Experimental results for standard optimization benchmarks indicate that the proposed algorithm rather than pure CSO can improve performance on finding the best global solution and achieves better accuracy level of convergence in less iteration.

II. OVERVIEW OF DIABETES CLASSIFICATION:

Classification is one of the most important decision making techniques in many real world problems. In this paper, the main purpose is to classify the data as diabetic or non-diabetic and improve the classification accuracy. For many classification problems, the higher number of samples chosen, but it doesn’t lead to higher classification accuracy.

In many cases, the performance of the algorithm is high in the perspective of speed, but the accuracy of data classification is low. Diabetes dataset taken from the UCI repository dataset for evaluate the proposed CSODT

R. Pavithra
Research scholar
Dept. of Computer science
Periyar University, Salem-11
rspavithrausa@gmail.com

H. Hannah Inbarani
Assistant Professor
Dept. of Computer science
Periyar University, Salem-11
hhinba@gmail.com
The diabetes classification method projected in the fig.1.

**Figure 1: Classification method**

The main task of this research work is partitioned into three stages:

(i) First one is Classification accuracy, achieve with individual model.

(ii) Second one is, Ensemble model used to achieve high accuracy compare to its individual model.

(iii) Finally the Third one is, Feature Selection technique applied on best ensemble model in order to achieve high accuracy which is computationally efficient.

In this work, we have used various classification techniques for classification of diabetic and non diabetic data. This data set is a binary class problem (i.e in the class attribute is zero’s and one’s) data set either is diabetic or non diabetic.

### III . PROPOSED METHODS FOR CLASSIFICATION:

We describe the CSO algorithm and the decision tree classifier algorithms . We compare proposed algorithm to various classifiers like as PSO,SOM,J48,PCA.

#### A. Decision tree classifier:

Decision tree induction is the learning of decision trees from class labeled training tuples. A decision tree is a flowchart-like tree structure, where each internal node (non leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The top most node in a tree is the root node.

The decision tree algorithm is well known for its robustness and learning efficiency with a learning time complexity of $O(\log 2n)$. C4.5 has been listed in the top 10 algorithms in data mining [55]. It is a popular statistical classifier developed by RossQuinlan in 1993. Basically, C4.5 is an extension of Quinlan’s earlier ID3 algorithm. In C4.5 the Information Gain split criterion is replaced by an Information Gain Ratio criterion which penalizes variables with many states. C4.5 can be used to generate a decision tree for classification. The learning algorithm applies a divide and conquer strategy [40] to construct the tree. The sets of instances are accompanied by a set of cats (attributes). This classifier has additional features, such as handling missing values, categorizing continuous attributes, pruning decision trees, deriving rules, and others.

### B. Cat Swarm Optimization

Cat Swarm Optimization is a new optimization algorithm in the field of swarm intelligence [4]. The CSO algorithm models the behavior of cats into two modes: ‘Seeking mode’ and ‘Tracing mode’.

Swarm is made of initial population composed of particles to search in the solution space. For example, we can simulate birds, ants and bees and create Particle swarm optimization, Ant colony optimization and Bee colony optimization respectively. Here, in CSO, we use cats as particles for solving the problems.

In CSO, every cat has its own position composed of D dimensions, velocities for each dimension, a fitness value, which represents the accommodation of the cat to the fitness function, and a flag to identify whether the cat is in seeking mode or tracing mode.

The final solution would be the best position of one of the cats. The CSO keeps the best solution until it reaches the end of the iterations [5].

Cat Swarm Optimization algorithm has two modes in order to solve the problems which are described below:

#### a. Seeking Mode

For modeling the behavior of cats in resting time and being-alert, we use the seeking mode. This mode is a time for thinking and deciding about next move. This mode has four main parameters which are mentioned as follow: seeking memory pool (SMP), seeking range of the selected dimension (SRD), counts of dimension to change (CDC) and self-position consideration (SPC). The process of seeking mode is describes as follow:

Step1: Make $j$ copies of the present position of cat$k$, where $j = SMP$. If the value of SPC is true, let $j = (SMP - 1)$, then retain the present position as one of the candidates.
Step2: For each copy, according to CDC, randomly plus or minus SRD percent the present values and replace the old ones.

Step3: Calculate the fitness values (FS) of all candidate points.

Step4: If all FS are not exactly equal, calculate the selecting probability of each candidate point by (1), otherwise set all the selecting probability of each candidate point be 1.

Step5: Randomly pick the point to move to from the candidate points, and replace the position of catk.

If the goal of the fitness function is to find the minimum solution, FSb = FSmax, otherwise FSb = FSmin

b. Tracing Mode

Tracing mode is the second mode of algorithm. In this mode, cats desire to trace targets and foods. The process of tracing mode can be described as follow:

Step1: Update the velocities for every dimension according to (2).

Step2: Check if the velocities are in the range of maximum velocity. In case the new velocity is over-range, it is set equal to the limit.

\[
V_{k,d} = V_{k,d} + c_1 r_1 (X_{best.d} - X_{k,d})
\]

Step 3: Update the position of cat k according to (3).

\[
X_{k,d} = X_{k,d} + V_{k,d}
\]

Xbest.d is the position of the cat, who has the best fitness value, Xk.d is the position of cat k, c1 is an acceleration coefficient for extending the velocity of the cat to move in the solution space and usually is equal to 2.05 and r1 is a random value uniformly generated in the range of [0,1]

V. EXPERIMENTAL WORK

In this research work, we have used MATLAB code and WEKA tool, we have used various individuals and hybrid classification models for classification of diabetes data. The analysis of models is done in two steps: first model is training and the second one is testing.In this research work using various data mining techniques like, PSO, SMO (Sequential Minimal Optimization), J48, PCA and are trained using a randomly training data set and after that the testing of the trained models is done using randomly tested data set. Partitions of data plays a very important task in accuracy of models.

Below figure shows the fitness function of the cat swarm optimization. Figure 1 shows that the proposed CSO can find the best solution of cat in less iteration. As shown, proposed CSO can get the optimal solution in 126th iteration. Mean while, the best solution of pure CSO is 0.004 in the 3500th iteration.

As we shown in Figure 3 the optimal solution of axis parallel function is zero and both algorithms can find it. But the proposed CSO, in comparison with pure CSO can find it in less iteration.

Fig. 2: Experimental Results of fitness function
VI. RESULT AND CONCLUSION

In the optimization theory, many algorithms are proposed and some of them are based on swarm intelligence. These types of algorithms imitate the behavior of animals. Cat Swarm Optimization (CSO) is a new optimization algorithm for finding the best global solution of a function which imitates the behavior of cats and models into two modes. In comparison with Particle Swarm Optimization CSO can find the better solution. In this research different classification algorithm compared with CSODT, according to the experimental result the proposed model CSODT provides higher accuracy 95.34% compared to other algorithms. Further enhancement of the research to includes various Evolutionary Optimization techniques, Genetic algorithm, classification, association rules etc.

REFERENCES


Table 1: RESULTS OF DIFFERENT CLASSIFICATION ACCURACY ALGORITHMS

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Types</th>
<th>Correctly Classified Instances</th>
<th>Incorrectly Classified Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSODT</td>
<td>95.34%</td>
<td>04.65%</td>
<td></td>
</tr>
<tr>
<td>PSO</td>
<td>80.51%</td>
<td>19.48%</td>
<td></td>
</tr>
<tr>
<td>SOM</td>
<td>81.15%</td>
<td>18.84%</td>
<td></td>
</tr>
<tr>
<td>J48</td>
<td>77.34%</td>
<td>22.65%</td>
<td></td>
</tr>
<tr>
<td>PCA</td>
<td>75.52%</td>
<td>24.47%</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3: Experimental Results of proposed CSODT

Fig. 4: Experimental Results for various algorithms


