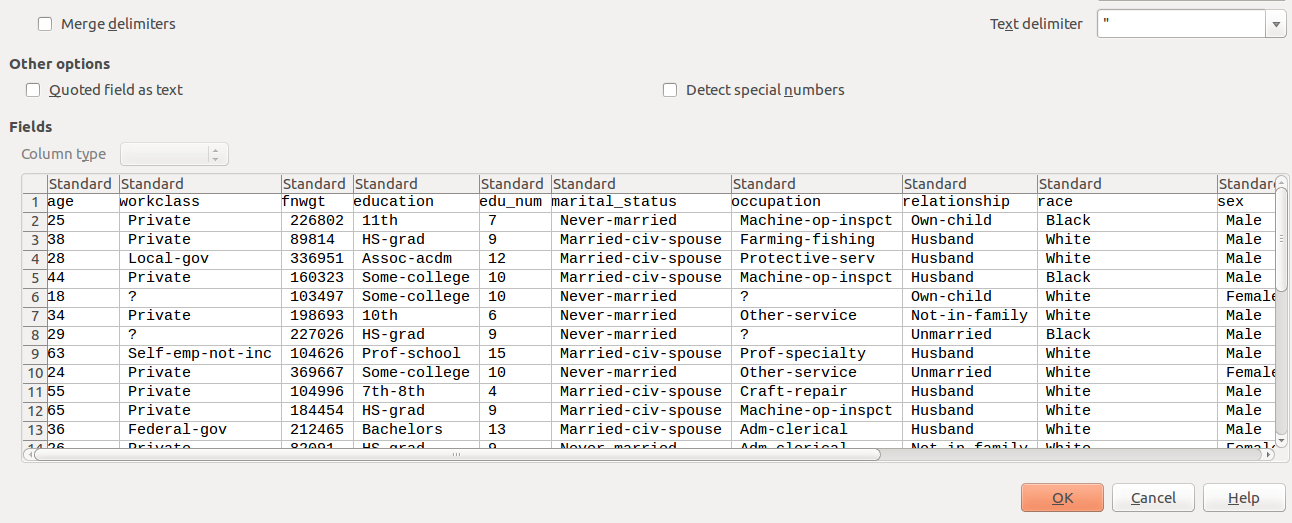
**RESULTS AND ANALYSIS**

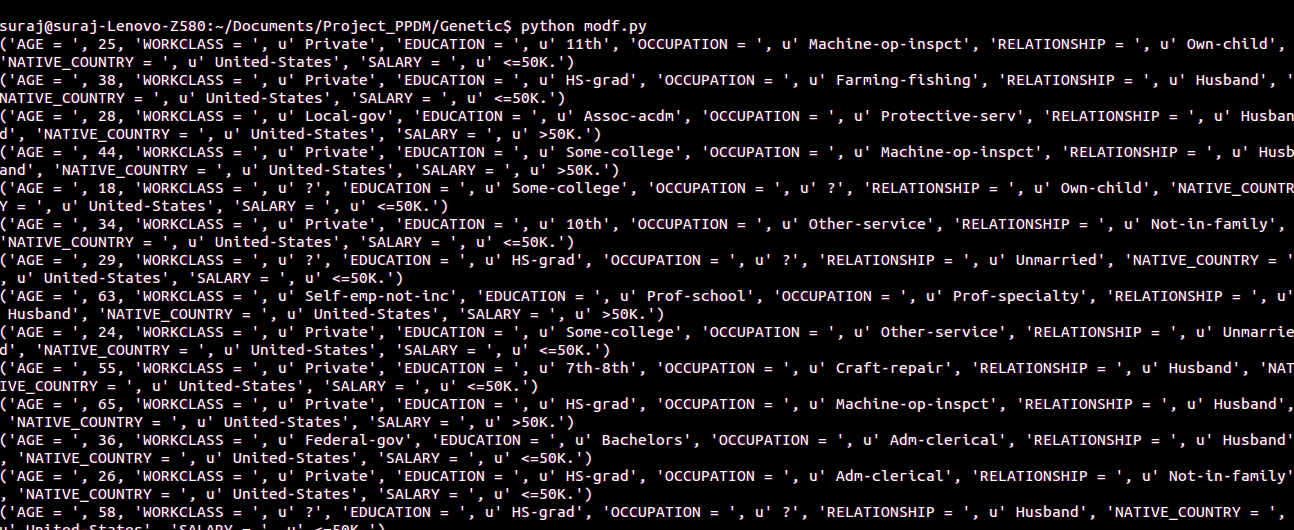
We adopt the adult dataset from the UCI machine learning repository for the experiments. The data set is then subjected to feature set partitioning and selection initially using the DMPD algorithm with genetic algorithm based search. The k value was initially set to three. The initial population is randomly generated as a sequence of bits equal to the number of quasi identifiers. The fitness value function uses a probabilistic sigmoid function to assign values to each population for comparison. The population is then subjected to the genetic algorithm search optimization technique and converged to a minimum threshold of fitness value difference.

The same data set was further subjected to feature selection using cuckoo search algorithm under k – anonymity value constraints. The initial population is randomly generated as a sequence of bits equal to the number of quasi identifiers. This population is then subjected to a random walk to generate closest set of values as compared to the initial population. Fitness values are generated for both parent and random populations generated with a probabilistic replacement of the parent population with the randomly generated population. The algorithm is then converged to the same minimum threshold of fitness value difference is successive iterations.

The final selected feature set are then subjected to anonymization techniques of generalization and suppression techniques to obtain the final data set.



**Figure 10:** UCI Adult Repository Dataset



**Figure 11:** Table of Data

**GA based DMPD Performance**

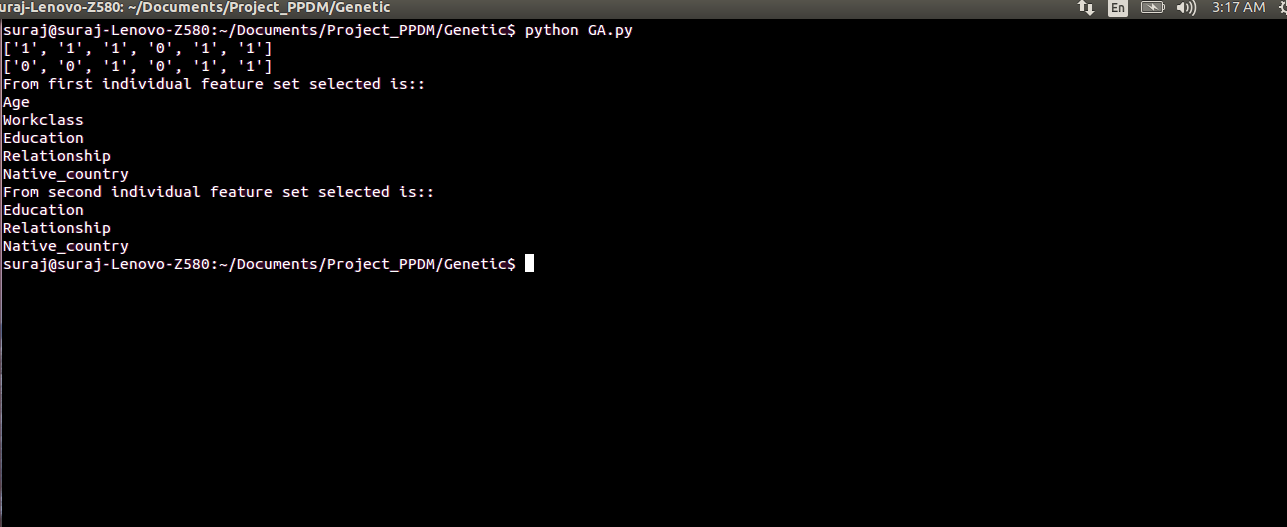
DMPD uses a general framework of feature set partitioning that can be applied prior to transforming feature values. In other words, DMPD can wrap any existing approach for anonymizing datasets and capture it for anonymizing projections following evaluation of partitioning fitness.

The DMPD is capable of dealing with different k-anonymity constraints and inducers without any significant effect on classification accuracy.

When compared to the state-of-the-art k-anonymity methods, DMPD classifiers provide an equivalent or slightly higher degree of accuracy.

DMPD, unlike other methods, does not use any prior knowledge. In TDS, TDR, Incognito and GA-based algorithms, the user is required to provide a taxonomy tree for categorical features. This makes these methods difficult to implement.

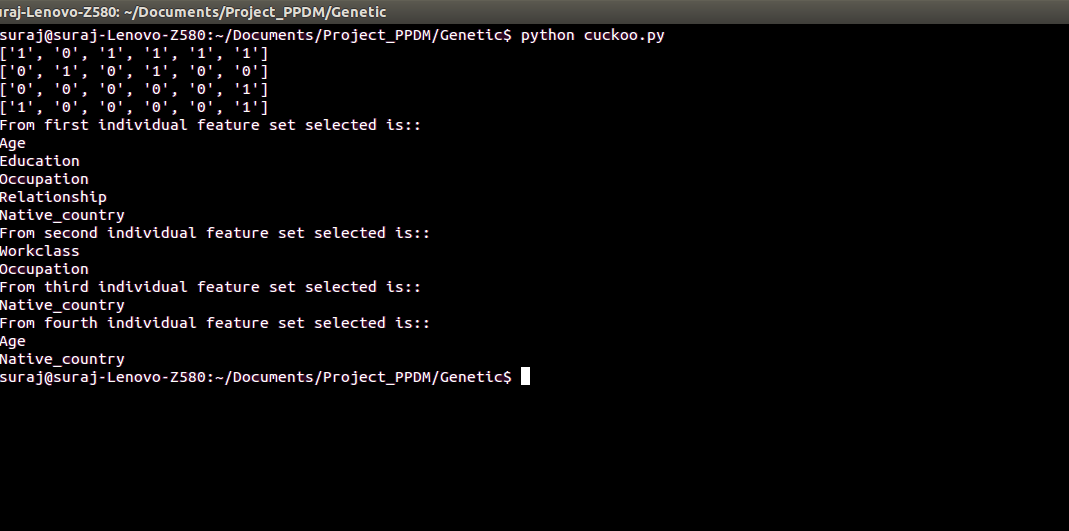
DMPD provide results without using any generalization of categorical feature values. Very possibly this feature results in deeper knowledge than the state-of-art generalization approaches.

DMPD can provide useful information regarding the tradeoff between k-anonymity constraints and classification performance.

**Figure 12:** Output Snippet of GA based DMPD Implementation.

**Performance of CS Algorithm**

The proposed algorithm has been validated and compared with other algorithms such as genetic algorithms and particle swarm optimization. Simulations and comparison show that CS is superior to these existing algorithms for multimodal objective functions. This is partly due to the fact that there are fewer parameters to be fine-tuned in CS than genetic algorithms. Furthermore, our simulations also indicate that the convergence rate is also quicker. This also means that we do not have to fine tune these parameters for a specific problem.

Subsequently, CS is more generic and robust for many optimization problems, comparing with other metaheuristic algorithms. This potentially powerful optimization strategy can easily be extended to study multiobjecitve optimization applications with various constraints, even to NP-hard problems. Further studies can focus on the sensitivity and parameter studies and their possible relationships with the convergence rate of the algorithm. Hybridization with other popular algorithms such as PSO will also be potentially fruitful.

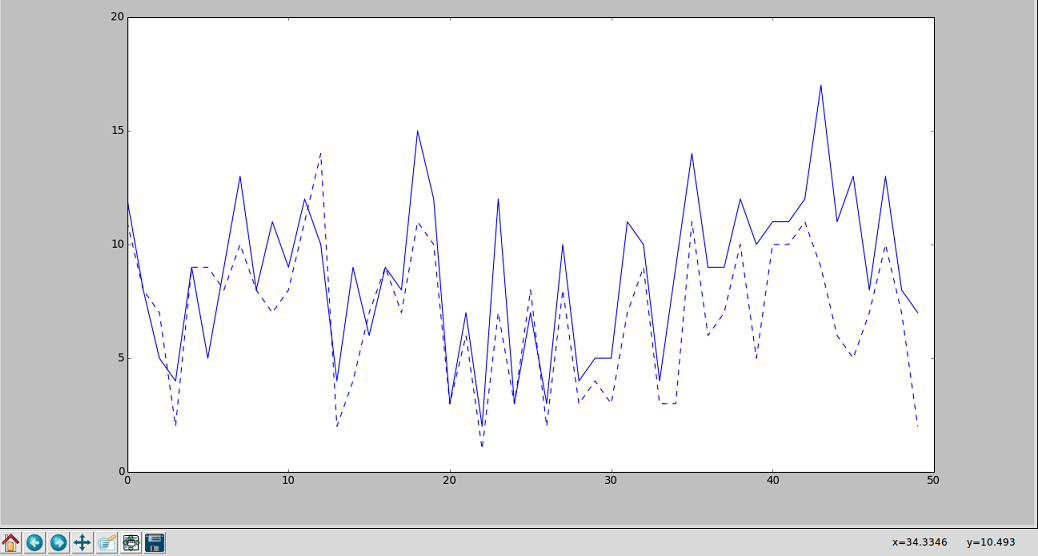
**Figure 13:** Output Snippet of CS based Implementation.

**Comparative Study**

The performance comparison of feature set partitioning and selection using both genetic algorithm (DMPD) and cuckoo search is carried out as follows.

The minimum fitness value difference between two successive iterations which forms the stopping condition was kept a constant for both the algorithms. Both the algorithms were run on the same dataset for the same number of times (in this case 50). Since both the algorithms select the next generation randomly, the output of each run is a probabilistic combination of features used. The k – anonymity constraint was also kept constant for both algorithms (k = 3).

For each algorithm the number of iterations it takes to converge were recorded and plotted against the number of runs of each algorithm.



**Figure 14:** Comparative study of both Cuckoo Search and Genetic Algorithms.

Note: Dotted line indicates number of iterations taken by Cuckoo search to converge.

Solid line indicates number of iterations taken by GA to converge.

X – Axis indicates the number of runs of both algorithms.

From the graph plot it is evident that the cuckoo search algorithm generally converges faster than the genetic algorithm based search.

**CONCLUSION**

Given a database both the feature set selection algorithms used in this report optimally select the set of features that satisfy the k – value constraints. The cuckoo search algorithm converges to a final set of values faster than the genetic algorithm based search owing to the method of generating the next closest population.

A new data anonymization method by randomizing the data records is introduced in this project. We arbitrarily replace part of the values in each record while providing the data relationships in the entire data set, which is dissimilar from most anonymization methods. This method not only accomplishes a higher level of privacy protection, but also preserves more knowledge than the other anonymization methods. We can say that, the valuable relations which are less susceptible can be exposed more precisely than the sensitive ones.