SOFTWARE FAULT PREDICTION USING NATURE INSPIRED HYBRID CLUSTERING ALGORITHMS

2019

ABSTRACT
Faults are inevitable in any software. A software fault is a structural imperfection in a software system which may eventually lead to system failure. Due to constraints and high complexity involved in the software development process, it is difficult to develop and produce software without faults. If at a later stage a fault is detected, the cost and effort required to correct the same is high, in the software development process. As prevention is better than cure, faults predicted and removed during initial stages, could reduce the cost of software development and maintenance. Software fault prediction is thus useful for finding faults before testing the software. There are various techniques to predict software faults such as using prediction models based on statistical and machine learning methods, using software metrics, process metrics, and complexity metrics, etc. Sometimes, in different environments and with different methods, software metrics give varied results.

Data mining techniques such as clustering and classification are also used for fault prediction. When fault labels for modules are unavailable, clustering is one of the best-suited techniques. k-means is a popular numerical, unsupervised, non-deterministic and iterative clustering algorithm, which is suitable for producing good quality clusters. Researchers have widely used k-means and its hybrid algorithms for fault detection and finding the fault-proneness of a software module.

In this work, k-means++, a hybrid k-means algorithm, which postulates a technique to initialize the cluster centers before proceeding with the normal k-means algorithm, is implemented and applied for fault prediction. k-means++ algorithm has lower computation time, improved speed and lesser errors as compared to other hybrid clustering algorithms.

Nature-inspired algorithms which draw inspiration from nature and mimic the behavior of living creatures such as Firefly Algorithm, Beehive Algorithm, etc. are widely used to optimize clustering algorithms. The Bat Algorithm, which swarms with varying speeds, the Cuckoo Algorithm, which does not swarm but iterates with fitness selection improvement and Krill Herd Algorithm, where the group of krills together find optimal solution; are the three nature-inspired techniques, used here for optimizing the k-means++ algorithm to enhance clustering quality and to generate new hybrids of unprecedented performance.
Henceforth, the motivation for this work is to create hybrid k-means algorithm and optimize it using hybrid swarm based techniques which promise to give better convergence, lesser cost and more accuracy than other optimization techniques to improve software fault prediction. k-means++ algorithm is thus optimized using Cuckoo, Bat and Krill Herd Algorithms to generate quality clusters which consume lesser CPU time and produces accurate fault prediction results.

Five new algorithms viz CuckooK++ (using Cuckoo and k-means++ algorithm), KBat++ (using Bat and k-means++ algorithm), KKH++ (using Krill Herd and k-means++ algorithm), CKBat++ (using Cuckoo and Bat Algorithm on k-means++ algorithm) and CKKH++ (using Cuckoo and Krill Herd Algorithm on k-means++ algorithm) are designed in this work and implemented for software fault prediction to generate faulty and non-faulty clusters in software modules.

Experiments are conducted to compare the performance of the proposed algorithms with the existing hybrid clustering algorithms on the basis of various performance parameters such as accuracy, specificity, precision, sensitivity, recall, f-measure, CPU Time, g-mean, inter-cluster and intra-cluster distances, etc. The results indicate the efficient performance of Cuckoo, Bat and Krill Herd based k-means++ designed algorithms.

The hybrid algorithms generated in this work have good convergence but cannot be considered optimal. In the future, optimization of hybrid clustering algorithms can be performed using other nature-inspired techniques on various hybrid clustering algorithms. The experiments performed in this work are conducted in a controlled environment, so the work can be performed a step further and implemented in the real-time environment. Also, the designed algorithms can be upgraded further and can be implemented in big data applications.

**KEYWORDS:** Software Faults, Nature Inspired Algorithms, Software Fault Prediction, Clustering, k-means Algorithm, k-means++ Algorithm, Cuckoo Algorithm, Bat Algorithm, Krill Herd Algorithm