ABSTRACT

With the enormous amount of data stored in files, databases, and other repositories, it is important to develop a powerful analysis tool to extract interesting knowledge from data that could help in decision making. Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data. Data mining is defined as “The nontrivial extraction of implicit, previously unknown, and potentially useful information from data”.

Rule induction is an area of machine learning in which formal rules are extracted from a set of observations. The performance of the classification methods are estimated and evaluated according to predictive accuracy, speed, robustness, scalability, and interpretability. This research concentrates on the performance enhancement of rule induction algorithm through hybridizing the swarm intelligence techniques and implements it by using cooperative coevolution framework.

Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO) are the two important swarm intelligence techniques used in data mining. From the literature review, it is observed that the existing ACO based rule induction algorithms support only nominal attributes. It uses discretization technique to convert continuous attributes into nominal attributes. PSO based rule induction algorithms support only continuous attributes and use indexing techniques to represent nominal attributes.
Evolutionary algorithms require an encoding scheme to represent the attributes. Tabu search is a powerful stochastic optimization technique. The incorporation of tabu search (TS) as a local improvement procedure helps to explore the search space efficiently. Sequential based approach algorithm takes more time to discover a rule set from a large dataset. Cooperative coevolution framework is used to solve large scale problems.

This research is one such attempt that exploits the natural and biological phenomenon to bring about performance amelioration in rule induction algorithm. The main goal of this research work is to develop hybrid rule induction algorithms that support mixed attributes.

A rule induction algorithm was developed by using ACO for discovering unordered rule set for mixed attributes. Mixed kernel Probability Density Function (PDF) was employed to initialize and update the pheromone value for continuous attributes. Laplace confidence was used as a heuristic function. The quality of a rule was measured by using rule confidence threshold value. Two different random transition rules were employed to exploit and explore the search space efficiently and found the optimal exploring rate for each data set. C4.5 discretization process was used to find the ranges in the continuous domain. The proposed algorithm produced more accurate results for mixed attributes data set than the existing algorithms. In order to handle continuous attributes without discretization, Particle Swarm Intelligence technique was applied.
A hybrid rule induction algorithm was developed by coalescing the merit of ACO, PSO, and TS techniques. This hybrid algorithm consists of two phases and it is executed in sequential order. In the first phase, a nominal rule was created by using the pheromone concept and in the second phase the continuous attributes were added with nominal rule by using the excellence of PSO and TS. TS technique helps to improve the search capability of PSO. The hybrid algorithm produced a good result for mixed attributes than that of the pure ACO algorithm.

Also, a coevolution rule induction algorithm was developed to improve the performance and scalability of rule induction algorithm. It engaged two swarms simultaneously to find the solution in search space and it is executed in parallel. It produced more accurate result and a simple rule set when the data set had more number of attributes. The coevolution approach reduced the elapsed time as well as generated more accurate and simple rule set to make a good decision.

From the performance of the algorithms, it was found that the ACO based rule induction algorithm is suitable for nominal data sets while the sequential based algorithm is compatible for small mixed attribute data sets. The coevolution based rule induction algorithm is suitable for large data sets with a large number of mixed attributes.