CSCI 447 — Machine Learning: Soft Computing

Project #4

Assigned: November 13, 2017
Design Document Due: November 27, 2017
Project Due: December 11, 2017

For this final assignment, we are going to investigate the performance of several swarm-based algorithms on a data mining task. As such, you will need to implement two data mining algorithms, which will be described here – k-means clustering and DB-Scan. You will also implement a competitive learning neural network as an alternative "clustering" method. These will serve as baselines. Finally, you will implement a particle swarm optimization algorithm and an ant-colony optimization algorithm to cluster data. In all cases, we assume partition-based or density-based clustering and do not consider the related problem of hierarchical clustering.

K-Means Clustering

The k-means algorithm is a method of clustering data into k clusters, based on an assumption that the data can be partitioned and tend to follow a Gaussian model. Even if the data do not support these assumptions, k-means is simple to implement and often performs well. An algorithm for k-means is shown here.

Algorithm 1 K-Means Clustering

```
function KMeans($\mathcal{D}, k$)
    initialize $\mu_1, \ldots, \mu_k$ randomly
    repeat
        for all $\mathbf{x}_i \in \mathcal{D}$ do
            $c \leftarrow \arg\min_{\mu_j} d(\mathbf{x}_i, \mu_j)$        ▷ $d()$ is the distance between $\mathbf{x}_i$ and $\mu_j$.
            assign $\mathbf{x}_i$ to the cluster $c$
        end for
        recalculate all $\mu_j$ based on new clusters
    until no change in $\mu_1, \ldots, \mu_k$
    return $\mu_1, \ldots, \mu_k$
end function
```

DB-Scan

An alternative approach to clustering examines the relative densities of the data. This approach has the nice property that there is no assumption that the data are Gaussian, nor is there an assumption that the clusters partition "cleanly." For the DB-Scan method of clustering, we categorize the points in the data set into one of three categories: Core, Border, and Noise. Let θ denote a threshold on distance and $MinPts$ reflect a minimum number of points that must fall within some region. A Core point is one where there are at least $MinPts$ points within θ of that point. A Border point is one that falls within θ distance of some Core point. A Noise point is anything left over. An Algorithm for DB-Scan is shown here, assuming all points have already been categorized.

Algorithm 2 DB-Scan

```
function DB-SCAN($\mathcal{D}$)
    $currClustLbl \leftarrow 1$
    for all $p \in Core$ do do
        if $clustLbl[p] =$ “Unknown” then
            $currClustLbl \leftarrow currClustLbl + 1$
            $clustLbl[p] \leftarrow currClustLbl$
        end if
        for all $p' \in \theta$-neighborhood do
            if $clustLbl[p'] =$ “Unknown” then
                $clustLbl[p'] \leftarrow currClustLbl$
            end if
        end for
    end for
    return $clustLbl$
end function
```

Your assignment consists of the following steps:

1. Prepare a design document addressing the design of the various clustering algorithms. Be sure to include an explanation of your experimental design as well.

2. Select five (5) different data sets from the UCI KDD Repository (http://kdd.ics.uci.edu/) or the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/index.html). It is strongly recommended that you select data sets with common feature types and no missing values.

3. Implement k-means, DB-Scan, Competitive Learning, ACO-based clustering, and PSO-based clustering. You will need to investigate methods for ACO-based clustering and PSO-based clustering since those are not covered explicitly in class or in your readings.

4. Devise one or more appropriate performance measures for evaluating the results of clustering/unsupervised learning. You are free to consult the clustering and data mining literature for help, but be sure to cite your sources.

5. Develop a hypothesis focusing on convergence rate and final performance of the approach(es) used to solve the problem

6. Design an experiment to test your hypothesis, ensuring that the experiment yields sufficient data from which you can draw meaningful statistical conclusions.

7. Run experiments to test your hypothesis on the various methods that have been implemented.

8. Write a paper that incorporates the following elements, summarizing the results of your experiments:

    (a) Title and author name

    (b) A brief, one paragraph abstract summarizing the results of the experiments

    (c) Problem statement, including hypothesis

    (d) Description of algorithms implemented

    (e) Description of your experimental approach

    (f) Presentation of the results of your experiments

    (g) A discussion of the behavior of your algorithms, combined with any conclusions you can draw

    (h) Summary

    (i) References (you should have at least one reference related to each of the algorithms implemented, a reference to the data sources, and any other references you consider to be relevant)

9. Submit your design document, fully documented code, sample runs of each algorithm, and your paper.