Mining Spatial Data & Enhancing Classification Using Bio-Inspired Approaches

Poonam Kataria¹, Navpreet Rupal²

¹, ²Department of CSE, SUSCET, Tangori, Distt. Mohali, Punjab, India

Abstract: Data-Mining (DM) has become one of the most valuable tools for extracting and manipulating data and for establishing patterns in order to produce useful information for decision-making. It is a generic term that is used to find hidden patterns of data (tabular, spatial, temporal, spatio-temporal etc.). Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases. Extracting interesting and useful patterns from spatial datasets is more difficult than extracting the corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationship and spatial autocorrelation. Spatial data are the data related to objects that occupy space. A spatial database stores spatial objects represented by spatial data types and spatial relationship among such objects. Clustering is the process of partitioning a set of data objects into subsets such that the data elements in a cluster are similar to one another and different from the element of other cluster. The set of cluster resulting from a cluster analysis can be referred to as a clustering. Spatial clustering is a process of grouping a set of spatial objects into clusters so that objects within a cluster have a high similarity in comparison to one another, but are dissimilar to objects in other clusters. In this paper, enhancement of classification scheme is done by using various Honeybee Optimization and Firefly Optimization. There are numerous artificial intelligence techniques which help in data mining to get the optimized result of the query. Hybrid of K-Mean & Ward’s Method, Honeybee Optimization and Firefly Optimization will be compared on the basis of performance parameters of classification (precision, recall, cohesion, variance, F-Measure, H-Measure) and therefore enhancement will be done.

Keywords: Spatial Data Mining; Clustering; FFO; HBO; Hybrid K-Mean; Ward Method.

1. Introduction

Recent development in science and technology has given a big rise to the data in the data warehouse, so it becomes a cumbersome task to the information required. To solve this problem, various data mining techniques have been proposed. Data mining simply means to extract the data from the database, but to improve the result of query these data mining techniques have to be more efficient in order to get the optimized result of a query. Thus data mining is a process through which data is discovered with respect to its pattern and interrelationship because of which it has become a powerful tool. The process of data mining is shown in the figure 1.

![Figure 1: Process of data mining](image)

The various forms of data mining are as below:

a) **Spatiotemporal Data Mining:** Spatiotemporal data are data that relate to both space and time. It refers to the process of discovering patterns and knowledge from spatiotemporal data.

b) **Multimedia Data Mining:** It is discovery of interesting patterns from multimedia databases that store and manage large collections of multimedia objects, including image data, video data, audio data.

c) **Web Mining:** It is the application of data mining techniques to discover patterns, structures and knowledge from the web.

d) **Spatial data mining:** Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases [13]. Extracting interesting and useful patterns from spatial datasets is more difficult than extracting the corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationship and spatial autocorrelation [24]. Spatial data are the data related to objects that occupy space. A spatial database stores spatial objects represented by spatial data types and spatial relationship among such objects.

2. Proposed Methodology

As discussed earlier, the development in computer technology has advanced the generation and consumption of data in our daily life. As a consequence, challenges such as growing data in data warehouse, it becomes a cumbersome task to extract the relevant information and to do so data mining techniques are used. Data Mining has become a powerful tool to extract hidden patterns of data and is gaining importance as it helps in decision making in all spheres of life.

There are numerous artificial intelligence techniques which help in data mining to get the optimized result of the query. Hybrid of K-Mean & Ward’s Method, Honeybee Optimization and Firefly Optimization will be compared on the basis of performance parameters of classification.
(precision, recall, cohesion, variance, F-Measure, H-Measure) and therefore enhancement will be done. The objective of the work carried out in this paper can be stated in following points:

- To create Hybrid algorithm of K-Means and Ward’s Method
- Optimization using Honey-Bee and firefly algorithm
- To make enhancement through various performance parameters for evaluation of classification scheme.

3. Hybrid Approach Based On K-Mean and Ward’s Algorithm

Hybrid approach works as: first we select no. of cluster and generate K- homogenous cluster. After getting K cluster by applying k-mean, now find maximum value in the cluster by taking random threshold value. Check approaching value of cluster selected is nearby to maximum value or not. Now make clusters according to nearby value to maximum value of the cluster. Cluster are allocated within the range of max and min value of the previous cluster and in the end after repeating this at Last efficient and optimum cluster obtain.

**Algorithm:**

1. Set the no. Of cluster in k-means.
2. Apply k-means algorithm to generate k- homogenous Cluster.
3. After getting K-cluster applying ward’s algorithm.
4. Select all cluster as input to wards.
5. Take random threshold value as find distance as find maximum values in the clusters.
6. Check approaching value of cluster selected is nearby to max value or not?
7. Define cluster according nearby value or approaching value to max value of the previous cluster.
8. Cluster are allocated within the range of max and min value of the previous cluster.
9. Last efficient and optimum cluster obtain.

4. Optimization Based Clustering

This work deals with the implementation of the clustering by using different optimization algorithms that are Hybrid of K-mean and Ward’s method, Honey Bee optimization and Firefly Optimization then we compare the results of clustering and find the best optimization algorithm with high percentage of accuracy.

4.1 Honey Bee Optimization

The Bees Algorithm is a new population-based search algorithm, first developed in 2005 by Pham DT and Karaboga. D independently. The bees algorithm is a population-based search algorithm The Bees Algorithm is an optimisation algorithm inspired by the natural foraging behaviour of honey bees to find the optimal solution [6]. A honey-bee colony consists of queen(s) (best solution), drones (incumbent solutions), worker(s) (heuristic), and broods (trial solutions). A colony of honey bees can extend itself over long distances in multiple directions (more than 10 km). Flower patches with plentiful amounts of nectar or pollen that can be collected with less effort should be visited by more bees, whereas patches with less nectar or pollen should receive fewer bees. The bees who return to the hive, evaluate the different patches depending on certain quality threshold (measured as a combination of some elements, such as sugar content). They deposit their nectar or pollen go to the “dance floor” to perform a “waggle dance”. Bees communicate through this waggle dance which contains the following information: The direction of flower patches (angle between the sun and the patch), the distance from the hive (duration of the dance) and the quality rating (fitness) (frequency of the dance). This information helps the colony to send its bees precisely. Follower bees go after the dancer bee to the patch to gather food efficiently and quickly. The algorithm requires a number of parameters to be set:

1. Number of scout bees n,
2. Number of sites selected m out of n visited sites,
3. Number of best sites e out of m selected sites,
4. Number of bees recruited for best e sites nep or (n2),
5. Number of bees recruited for the other (m-e) selected sites which is nsp or (n1),
6. Initial size of patches ngh which includes site and its neighbourhood and stopping criterion,
7. Number of algorithm steps repetitions imax

4.2 Firefly Optimization

The firefly algorithm (FA) is a metaheuristic algorithm, inspired by the flashing behaviour of fireflies[30]. The primary purpose for a firefly’s flash is to act as a signal system to attract other fireflies. Xin-She Yang formulated this firefly algorithm by assuming:

- All fireflies are unisexual, so that one firefly will be attracted to all other fireflies; Attractiveness is proportional to their brightness, and for any two fireflies, the less bright one will be attracted by (and thus move to) the brighter one; however, the brightness can decrease as their distance increases; If there are no fireflies brighter than a given firefly, it will move randomly. The brightness should be associated with the objective function. Firefly algorithm is a nature-inspired metaheuristic optimization algorithm. Firefly Algorithm (FA) was first developed by Xin-She Yang in late 2007 and 2008 at Cambridge University, which
was based on the flashing patterns and behaviour of fireflies. In essence, FA uses the following three idealized rules:

- Fireflies are unisex so that one firefly will be attracted to other fireflies regardless of their sex.
- The attractiveness is proportional to the brightness, and they both decrease as their distance increases. Thus for any two flashing fireflies, the less brighter one will move towards the brighter one. If there is no brighter one than a particular firefly, it will move randomly.
- The brightness of a firefly is determined by the landscape of the objective function.

5. Experimental Work

The interface used for regionalization of spatial object is shown in Figure 3 GUI works as follows:

i) The main central GUI is linked to three windows.
   a. Select A Dataset
   b. Hybrid K-Mean And Ward’s Algorithm
   c. Optimization Techniques

ii) On clicking on the button SELECT DATASET as shown in Figure 3.2 a new pop-up window will be opened as shown in Figure 3.3. User can select different spatial dataset according to the choice.

iii) On clicking on the button HYBRID K-MEAN AND WARD’S ALGORITHM, a new window will be opened as shown in Figure 6

iv) First we enter the no. of k-cluster in the box which is in front of label ENTER THE VALUE OF K as shown in Figure 7

v) User when press the button HYBRID the results after applying Hybrid K-mean and Ward’s algorithm for solving regionalization issue in spatial clustering on different parameters i.e. Cohesion, Variance, Precision, Recall, F-measure and H-measure are shown.
After calculating the value for hybrid algorithm, click on BACK button. When we click on BACK Button it shows main window.

After applying Hybrid K-mean and Ward’s algorithm for solving regionalization problem, now we apply different optimization techniques on the result of hybrid algorithm to improve the efficient of clustering spatial objects in Figure 11

After opening OPTIMIZATION TECHNIQUES window we have two options to get efficient and homogenous cluster for Regionalization i.e. using HBO Algorithm or FIREFLY Algorithm as shown in Figure 12. When we click on “OPTIMIZATION TECHNIQUES” the window shown in the Figure 13 is Open

When we press the button HONEYBEE OPTIMIZATION the results for solving regionalization issue in spatial clustering on different parameters i.e. Cohesion, Variance, Precision, Recall, F-measure and H-measure are shown and also figure of Honeybee optimization and clusters comes after optimizing data.

When we press the button FIREFLY OPTIMIZATION the results for solving regionalization issue in spatial clustering on different parameters i.e. Cohesion, Variance, Precision, Recall, F-measure and H-measure are shown and also figure of Honeybee optimization and clusters comes after optimizing data.

B. Performance Parameters for Classification

Performance metric measures how well your data mining algorithm is performing on a given dataset. For example, if we apply a classification algorithm on a dataset, we first check to see how many of the data points were classified correctly. The various parameters taken for performance evaluation are as follows:
1) Precision:- It is also known as positive predictive value which is fraction of retrieved instances that are relevant
2) Recall :- It is defined as a set of relevant documents (e.g. the list of all documents on the internet that are relevant for a certain topic)
3) Cohesion:- It measures how closely objects in the same cluster are related
4) Variance:- Variance measures how distinct or well separated are clusters from each other

6. Results

Cohesion
The figure 15 indicate that FFO is highly cohesive as compared to Hybrid K. Means and Honey bee Optimization.

![Cohesion Graph](image)

Variance
As per the definition the variance should be close to 0, in the figure 4.2, FFO has less variance value i.e. 0.41 as compared to HBO and Hybrid K-means and Ward with value of 0.51 and 0.66 respectively.

![Variance Graph](image)

Precision
The below graph show that Firefly Optimization and Honey Bee Optimization has high precision value i.e. 0.68 whereas Hybrid K-means and Ward Method has 0.60.

![Precision Graph](image)

Recall
The value of recall in Hybrid K-mean & Ward method is 0.55, in HBO is 0.64 and is high in FFO i.e. 0.73.

![Recall Graph](image)

F-Measure
The graph below in figure 19 shows the value of f-measure obtained using various techniques. In Hybrid K-mean & Ward method , HBO and FFO the value of F-Measure is 0.57, 0.65 and 0.70 respectively.

![F-Measure Graph](image)

H-Measure
The graph below shows the value of h-measure obtained using various techniques. In Hybrid K-mean & Ward method, HBO and FFO the value of F-Measure is 0.57, 0.65 and 0.71 respectively.

![H-Measure Graph](image)
Comparison of Evaluation Parameters

For the purpose of analysis, the comparison of different parameters such as cohesion, variance, precision, recall, H-Measure and F-Measure on spatial dataset has been calculated using Hybrid K-Means & Ward’s Method, HBO and FFO are tabulated

Table 1: Comparison of Hybrid K-Means & Ward’s method, HBO, FFO on the basis of Evaluation parameters

<table>
<thead>
<tr>
<th>S.No</th>
<th>Parameters</th>
<th>Hybrid K-Mean &amp; Ward’s method</th>
<th>Honey Bee Optimization</th>
<th>Firefly optimization</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cohesion</td>
<td>0.51</td>
<td>0.55</td>
<td>0.62</td>
</tr>
<tr>
<td>2</td>
<td>Variance</td>
<td>0.66</td>
<td>0.51</td>
<td>0.41</td>
</tr>
<tr>
<td>3</td>
<td>Precision</td>
<td>0.60</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>4</td>
<td>Recall</td>
<td>0.55</td>
<td>0.64</td>
<td>0.73</td>
</tr>
<tr>
<td>5</td>
<td>F-Measure</td>
<td>0.573</td>
<td>0.65</td>
<td>0.70</td>
</tr>
<tr>
<td>6</td>
<td>H-Measure</td>
<td>0.57</td>
<td>0.65</td>
<td>0.71</td>
</tr>
</tbody>
</table>

7. Conclusion

The paper presents regionalization based on Hybrid K-Means and Ward’s Clustering algorithm using different optimization technique i.e Honey Bee Optimization and Firefly algorithm. In the paper, three algorithms Hybrid K-Mean and Ward Method, HBO and FFO are implemented on a spatial dataset taken from UCI Machine Learning Repository. Using each algorithm, some performance parameters such as Cohesion, Variance, Precision, Recall, F-Measure and H-Measure are calculated. It can be concluded that, H-Measure, F-Measure, Cohesion, Recall, Precision on dataset is more in Firefly Algorithm as compared to HK-Mean ward method and Honey Bee Optimization, while Variance is less in FFO. As seen in this paper work, FFO has been implemented successfully over Hybrid K-Mean & Ward Algorithm and HBO.

8. Future Work

In the present work we have implemented FFO based classification successfully using spatial data set taken from UCI Repository of Machine Learning Databases. For future work, we can combine some other artificial intelligence algorithm to get more optimized result and can make enhancement using some other parameters also.


[31] Yang, Xin–She, and Xingshi He.(2013) "Firefly algorithm: recent advances and applications." International Journal of Swarm Intelligence 1, no. 1 36-50.