UNDERSAMPLING USING LEADER ALGORITHM

OBJECTIVES

The main objective of this project is to implement the **Leader Algorithm** for undersampling the majority class in an imbalanced dataset and make that as a balanced dataset, and also to improve the accuracy of the classification before and after the undersampling process.

PROBLEM DESCRIPTION

In a imbalanced dataset which consists of two classes namely major class and minor class. The major class contains more number of instances compared to the minor class instance. Leader algorithm is implemented on the majority class and the number of instances will be minimized to get a better accuracy.

PROCEDURE

1. Data Preprocessing
   - Load the ecoli dataset.
   - Handle the missing values in the dataset if there is any.
   - If any dataset row have Boolean values like **"yes or no"**, **"positive or negative" ,**encode those columns as 0s or 1s for better utilization.
2. Implementation of Algorithm
   - Implement the Leader Algorithm to the imbalanced dataset and count the number of major and minor class instances.
   - Separate the major class instances and make them into a separate csv file.

2.1 Pattern Sum Calculation

- Calculate the pattern sum for each row of the major class instance by adding its features except the **"class"** feature.
- Then sort the pattern sums in ascending order for better classification.

2.2 Cluster Formation

- Now each pattern sum is considered to be a cluster representative or leader of the cluster and it should be compared with all the other data.
- Set a threshold value manually according to the difference between the pattern sum.
- If the comparison value is higher than the threshold value then that data row must be added to a new cluster and it should made as the new cluster's leader.
- Likewise compare all the pattern sum with other cluster leaders and make it into a cluster according to the threshold value.

2.3 Balancing Dataset

- After completing the above steps take the undersampled majority class instances and combine it with the minority class instances and save it as a new csv file.

3. Accuracy Check

- After creating a balanced dataset file , classify the balanced dataset using KNN classifier and check its accuracy level.
- The imbalanced dataset should also be classified using KNN classifier and find out its accuracy level.
- During the KNN classification the data must be split into 80% for training and 20% for testing for both the balanced and imbalanced dataset.

EXPERIMENT RESULTS

Table 1 : Ecoli Dataset (Imbalanced).

Total Number of features	7
Total number of major class instances	143
Total number of minor class instances	77
Total number of instances in the dataset	220
Features	Mcg , Gvh , Lip , Chg , Aac , Alm1 , Alm2

Table 2 : Test Accuracy for both datasets.

Data / Dataset	**Imbalanced Dataset**	**Balanced Dataset**
Total testing data in major class	29	23
Number of correctly Predicted data	29	23
Total testing data in Minor class	16	16
Number of correctly Predicted data	11	13
Accuracy	86.88%	92.30%

CONCLUSION

The Leader Algorithm for undersampling the imbalanced dataset has successfully implemented and it has reduced the number of instances in the major class and made the dataset into a balanced one. It also increased the accuracy of the prediction of the balanced dataset compared to that of imbalanced dataset.