IMPLEMENTATION OF MACHINE LEARNING IN AIR LINE DATA

P. Rajasree[1], B. Vimala devi[2], S. Gomathi[3]

[1] I MSc Data Analytics, Department of Computer Science(PG)
PSGR Krishnammal College for Women, Coimbatore, India

[2] I MSc Data Analytics, Department of Computer Science(PG)
PSGR Krishnammal College for Women, Coimbatore, India

[3] Assistant Professor, Department of Computer Science(PG)
PSGR Krishnammal College for Women, Coimbatore, India.

Abstract:

The main objective of the paper is to implement classification and clustering algorithm such as Naive Bayes, K-Means, and Farthest First clustering algorithm using Weka to analyse airline dataset. In addition to that, performance analysis of these algorithms shows that Farthest First outperforms the other algorithms. Data set has been obtained from UCI machine learning repository. The algorithms has been discussed with the results obtained from the tool. The confusion matrix for classification algorithm and the cluster instances for clustering algorithm is also shown in the result.

Introduction:

Data mining is the method of exploring unknown data patterns according to different viewpoints for classification into beneficial info, which is composed and accumulated in common portions, such as data warehouses for effectual study, data mining algorithms, smoothing business decision making and other information necessities to eventually cut costs and upsurge revenue.

Data mining comes under the umbrella term of "business intelligence," and could be considered a form of BI[3]. Data mining is used to collect relevant information and gain insights. Moreover, business intelligence could also be thought as the result of data mining. As stated, business intelligence comprises using data to obtain insights. Data mining is the collection of essential data, which will eventually lead to solutions through in-depth analysis.

The relation between data mining and business intelligence can be thought of as a cause-and-effect relationship .Data mining pursuits for the "what" and business intelligence processes reveal the "how" and "why"[1]. Analysts make use of data mining to find the information they need and use business intelligence to define why it is important. Data mining is a practice used by companies to turn raw data into useful information.

Data mining is applied efficiently in various fields such as business atmosphere, weather forecast, medicine, transportation, healthcare, insurance, government…etc. Important elements such as learning more about their customers, to develop more effective marketing tactics, increase sales and decrease costs can be done through data mining which leads to business intelligence.

Data Mining Algorithm:

An *algorithm* in data mining is a set of heuristics and calculations that creates a model from data. For creating a model, the algorithm first analyses the data you offer, observing for particular types of patterns or trends.

The algorithm uses the outcomes of this analysis over many iterations to find the optimal factors for producing the mining model. These parameters are then applied across the whole data set to extract actionable patterns and complete statistics.

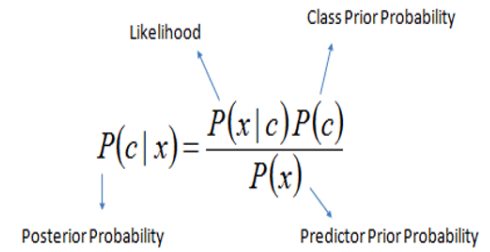
Various forms can be taken by the mining model that an algorithm creates from your data, including

- A set of clusters that describes how the cases in a dataset are associated.
- A decision tree that predicts a result, and describes how different conditions affect that outcome.
- A mathematical model that estimates sales.
- A set of rules that define how data can be grouped together in a transaction, and the prospects that products are purchased together.

1. Naive Bayes Classification Algorithm:

Naive Bayes is a statistical classification method based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, exact and consistent algorithm. Naive Bayes classifiers have great correctness and speed on huge datasets.

Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features. For example, a loan applicant is desirable or not depending on his/her income, previous loan and transaction hi story, age, and location. Even if these features are interdependent, these features are still considered independently. This assumption simplifies computation, and that's why it is considered as naive. This assumption is called class conditional independence.



$$P(c|X) = P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$$

Fig 1.1 Naïve Bayes Formula

2. K-Means Clustering Algorithm:

One of the simplest unsupervised learning algorithms is the K-means algorithm that solves the well-known clustering problem. The technique follows a simple and relaxed approach to classify a given data set through a certain number of clusters (assume k clusters) [2]. Defining k-centres, one for each cluster is the main idea of this algorithm.

These centres should be placed in a cunning way because different results are obtained from different locations. So, the healthier choice is to place them far away as much as possible from each other. The succeeding step is to take each point fitting to a given data set and associate it to the nearest center. When no point is awaiting, the first step is completed and an early group age is done. We need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step at this point.

After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been created. As a result of this loop we may notice that the k-centers alter their location step by step until no more changes are done or in other words centers do not move any more.

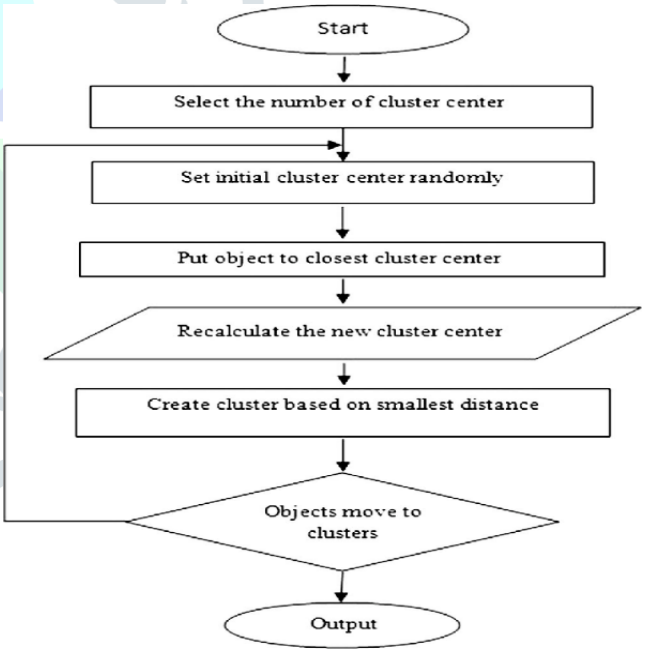


Fig 2.1 K-Means Algorithm Flowchart

3. Farthest First Clustering Algorithm:

Farthest first is a modified of K-Means that places each cluster center in turn at the point further most from the existing cluster center. This point must lies within the data area. [3]This significantly increases the clustering speed in most of the cases since less reassignment and adjustment is

needed. Farthest First Algorithm is suitable for large dataset but it creates non-uniform clusters.

The farthest-first traversal k-center (FFT) is a fast and greedy algorithm. In this algorithm k points are first selected as cluster centers. The first center is select randomly. The second center is greedily select as the point furthest from the first. Each remaining center is determined by greedily selecting the point farthest from the set of already chosen centers, and the remaining points are added to the cluster whose center is the closest.

Obtained Results:

The following figures A,B and C are the implementation results of the above three algorithms.

Naïve Bayes Algorithm:

From the following result, it is known that Naïve Bayes algorithm takes 0.8 seconds to build a model and classify the attributes for this given dataset.

```
Time taken to build model: 0.8 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      341194               63.2564 %
Incorrectly Classified Instances    198189               36.7436 %
Kappa statistic                          0.2399
Mean absolute error                      0.426
Root mean squared error                  0.4796
Relative absolute error                 86.2198 %
Root relative squared error             96.5044 %
Total Number of Instances           539383

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area
               0.757     0.522     0.643      0.757     0.696      0.245    0.668
               0.478     0.243     0.612      0.478     0.537      0.245    0.668
Weighted Avg.  0.633     0.398     0.630      0.633     0.625      0.245    0.668

=== Confusion Matrix ===

     a      b   <-- classified as
 226388  72731 |      a = 0
 125458 114806 |      b = 1
```

Fig A Naïve Bayes implementation result

Time taken to build model: 0.8 seconds

K-Means clustering algorithm:

From the following result ,it is known that K-Means clustering algorithm takes 1.27 seconds to cluster this given dataset.

```
Attribute      Full Data          0          1
                (539383.0) (247360.0) (292023.0)
=============================================
Airline               WN         WN         OO
Flight          2427.9286  1748.8759  3003.1247
AirportFrom           ATL        ATL        ATL
AirportTo             ATL        ATL        ATL
DayOfWeek               4          6          3
Time              802.729   837.3786   773.3788
Length            132.202   139.8814   125.6971
Delay                   0          1          0



Time taken to build model (full training data) : 1.27 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      247360 ( 46%)
1      292023 ( 54%)
```

Fig B Implementation Result of K-Means Algorithm

Time taken to build model:1.27 seconds

Farthest First Clustering Algorithm:

From the following result ,it is known that K-Means clustering algorithm takes 0.45 seconds to cluster this given dataset.

```
FarthestFirst
==============

Cluster centroids:

Cluster 0
        WN 915.0 CMH STL 2 1120.0 90.0 1
Cluster 1
        OO 7781.0 EKO SLC 6 320.0 72.0 0



Time taken to build model (full training data) : 0.45 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       301947 ( 56%)
1       237436 ( 44%)
```

Fig C Implementation result of Farthest First Algorithm

Time taken to build the model :0.45 seconds

Conclusion:

- Therefore it is examined that Naïve bayes algorithm takes 0.55 second to classify the attributes on this dataset and to create a confusion matrix.
- K-Means clustering algorithm takes 1.27 seconds to complete clustering of attributes on this dataset.

  Clustered Instances : 247360

  Non-clustered Instances : 292023

- Farthest First algorithm takes 0.45 seconds to complete the clustering of instances.

  Clustered Instances : 301947

  Non-clustered Instances : 237436

When comparing both the clustering algorithms , Farthest First Algorithm clusters more number of instances at a higher rate than K-Means clustering algorithm.

References:

[1] Nikam, Sagar S. "A comparative study of classification techniques in data mining algorithms." *Oriental journal of computer science & technology* 8.1 (2015): 13-19.

[2] Dhanachandra, Nameirakpam, Khumanthem Manglem, and Yambem Jina Chanu. "Image segmentation using K-means clustering algorithm and subtractive clustering algorithm." *Procedia Computer Science* 54 (2015): 764-771.

[3] Tan, Pang-Ning. *Introduction to data mining*. Pearson Education India, 2018.

[4] Zafar, Muhammad Husnain, and Muhammad Ilyas. "A clustering based study of classification algorithms." *International Journal of Database Theory and Application* 8.1 (2015): 11-22.