A Hybrid Implementation of K-Means and HAC Algorithm and Its Comparison with other Clustering Algorithms

Anita Ganpati, Jyoti Sharma

Abstract:- There is a huge amount of data which is being produced everyday in Information Technology industry but it is of no use until converted into useful information. Data mining is defined as the process of extracting of hidden predictive information from large databases. Data mining provides an easy and timesaving concept to extract the useful information from large database instead of going through the whole database. There are various data mining techniques and clustering is one of them. Clustering algorithms especially draws significant attention of researchers all around the world because it makes an easy availability of the same data in form of clusters. There are various types of clustering algorithms available in the literature, with each algorithm having its own pros and cons. In this research paper, a hybrid implementation of k-Means and HAC clustering algorithm is presented. Also, the hybrid approach is compared with four other clustering algorithm namely k-Means, DT, HAC, VARCHA. The hybrid implementation has been done using Python scripting language and SCIKIT LEARN open source tool was used for the performance comparison of the algorithms. The various parameters used for comparison were accuracy, precision, recall and f-score. The results show that the performance of hybrid algorithm is found to be quite better than the existing ones.

Keywords: Data Mining, Clustering, k-Means, DT, HAC, VARCHA, Python and SCIKIT.

I. INTRODUCTION

Nowadays, in all professional fields Data Mining is becoming a very attractive field. Data mining is often defined as finding hidden information in a database [7]. Data mining technology blends traditional data analysis methods with sophisticated algorithm for processing large volume of data. It helps in analysing and exploring new type of data and for analyzing old type of data in new ways. Data mining is integral part of knowledge discovery in databases. According to Bhavani, “Data Mining is the process of extracting useful information, from large quantities of data possibly stored in data bases” [4]. “Data Mining is an integral part of knowledge discovery in databases, which is the overall process of converting raw data into useful information”[8]. Clustering is an important data mining technique used to place data elements into related groups without advance knowledge of the group definition. Clustering divides data into meaningful or useful groups that is “Clusters” [14]. As shown in Figure 1, the clusters are obtained from a data set. The objects of similar properties are placed in one class of objects and a single access to the disk makes the entire class available.

Figure 1: Formation of Clusters in Clustering [15]

II. LITERATURE REVIEW

M.Sathyaa Deepa [6] has compared various clustering techniques and their characteristics. There are many methods to form clusters. The four important methods of clustering namely partition clustering, hierarchical clustering, density-based clustering and grid-based clustering are studied in detail. Jyoti et al.[5] in their paper presented a comparative study of an analysis of grid based clustering algorithms in data mining. They compared grid based clustering algorithms namely STING, CLIQUE, and WAVE CLUSTER. The algorithms were compared on the basis of the different parameters such as noise, shape, time complexity, data sets, accuracy and many more. After analysing the results it was found that the time complexity of all the three is same but the approaches are different. The shape of CLIQUE and WAVE CLUSTER are same as compared to STING. WAVE CLUSTER is not sensitive to noise which is very beneficial and the accuracy of wave cluster is good than clique and sting. Aastha Joshi [3] has compared five types of clustering techniques- k-Means clustering, Hierarchical clustering, DBSCAN clustering OPTICS, STING. The conclusion of comparison was that the K-mean algorithm has biggest advantage of clustering large data sets and its performance increases as number of clusters increases. But its use is limited to numeric values. Ravindra Jain proposed a hybrid clustering algorithm for data mining. The research focused on fast and accurate clustering. Its performance is compared with the traditional
k-Means & KHM algorithm. The result obtained was that the proposed hybrid algorithm is much better than the traditional k-means & KHM algorithm [9]. Tapas Kanungo et al. [13] discussed that in k-means clustering, a set of n data points in d-dimensional space $\mathbb{R}^d$ and an integer k is given and the problem is to establish a set of k points in $\mathbb{R}^d$, called centers, so as to minimize the mean squared distance from each data point to its nearest center. A popular heuristic for k-means clustering is Lloyd's algorithm. In their paper, they presented a simple and efficient implementation of Lloyd's k-means clustering algorithm, which they called the filtering algorithm. According to them, this algorithm is simple to implement, requiring a k-d tree as the only main data structure. S. Revathi et al. [12] performed a comparative study of clustering algorithms across two different data items. The performance of the various clustering algorithms is compared based on the time taken to form the estimated clusters. The experimental results of various clustering algorithms to form clusters are depicted as a graph. Thus it can be concluded as the time taken to form the clusters increases as the number of cluster increases. The simple k-Means takes the longest time to perform clustering. Manju Kaushik et al. in their paper compared k-Means clustering and hierarchical clustering techniques. They discussed the strengths and weaknesses of both clustering techniques in detail. The k-Mean algorithm has the big advantage of clustering large data sets and its performance increases as the number of clusters increases. According to them the performance of k-Mean algorithm is better than hierarchical clustering algorithm. When using huge dataset, k-Means algorithm is faster than other clustering algorithm and also produces quality clusters[10]. Bharat Chaudhari et al. analysed three major clustering algorithms i.e. k-Means, hierarchical clustering and density based clustering algorithm and compared the performance of these three major clustering algorithms on the aspect of correctly class wise cluster building ability of algorithm. They concluded that the performance of k-Means algorithm is better than hierarchical clustering algorithm. However, density based clustering algorithm is not suitable for data with high variance in density. Also, k -Means algorithm is produces quality clusters when using huge dataset[11].

III. OBJECTIVES

The main objective of the study is to propose a new clustering algorithm. However the specific objectives of the study are:

1. To have an understanding of data mining concepts and techniques.
2. To study the different clustering techniques and algorithms.
3. To propose a hybrid clustering algorithm.
4. To compare the hybrid algorithm with its other counterparts namely k-Means, DT, VARCHA and HAC.

IV. RESEARCH METHODOLOGY

The research methodology firstly followed a theoretical approach for the study of data mining, various clustering techniques and algorithms. The selection of techniques for the objectives includes literature survey, articles, books, research paper and internet. Then secondly, an empirical study is performed to evaluate clustering algorithms. The hybrid implementation has been done using Python scripting language. The SCIKIT-LEARN open source tool was used for comparison.

V. ANALYSIS OF RESULTS

The paper aims to develop a novel clustering algorithm which is a hybrid of HAC and k-Means clustering. The novel algorithm is implemented and compared for performance along with four other algorithms such as K-means, HAC clustering, VARCHA clustering and Decision Tree based clustering. The methodology firstly follows a theoretical approach for the study of data mining, various clustering techniques and its algorithm. The algorithms are implemented using Python as a language on a standard dataset. The tool utilized is SCI-KIT-LEARN which utilizes SCIPY and NUMPY tools at the backend. The proposed hybrid algorithm is designed such as to combine HAC clustering by first deciding hierarchy for various clusters and then applying K-means clustering based on distance of various points from centroids. The results for the various algorithms have been implemented and plotted using MS-EXCEL and the values of accuracy, precision, recall and f-score are compared. The computed values of Accuracy, Precision, Recall and F-measure using SCIKIT tool are shown in Table 1. As shown from the values in the table the proposed hybrid algorithm has the highest value i.e. 98.66 for accuracy parameter. Also the proposed algorithm is better in terms of precision than k-Means, DT, HAC and VARCHA.

Table 1: Computed Values of Accuracy, Precision, Recall and F-measure.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means</td>
<td>92.98</td>
<td>92.0</td>
<td>89.32</td>
<td>90.64</td>
</tr>
<tr>
<td>DT</td>
<td>94.09</td>
<td>88.99</td>
<td>96.03</td>
<td>92.38</td>
</tr>
<tr>
<td>HAC</td>
<td>93.72</td>
<td>86.99</td>
<td>98.6</td>
<td>92.01</td>
</tr>
<tr>
<td>VARCHA</td>
<td>86.71</td>
<td>72.89</td>
<td>91.76</td>
<td>81.24</td>
</tr>
<tr>
<td>PROPOSED HYBRID</td>
<td>98.66</td>
<td>100</td>
<td>94.94</td>
<td>97.40</td>
</tr>
</tbody>
</table>

It is evident from the results that the accuracy, precision and f-measure of the proposed algorithm is better than the other compared algorithms and recall value of HAC is more than other algorithm. So the overall performance of proposed hybrid algorithm is better than the other four algorithms. The Figure 2 is showing the visualization of the results on the basis of the values.
VI. CONCLUSION AND FUTURE SCOPE

In this research paper, a hybrid implementation of k-Means and HAC clustering algorithm is presented. Also, the hybrid approach is compared with four other clustering algorithm namely k-Means, DT, HAC, VARCHA. The various parameters used for comparison were accuracy, precision, recall and f-score. The results show that the performance of hybrid algorithm is found to be quite better than the existing ones.

REFERENCES