Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 06, June -2018

COMPARITIVE ANALYSIS OF MACHINE LEARNING ALGORITHM

Shweta Verma[1], Yamini Chouhan[2]

1Research Scholar, Department of Computer Science and Engineering, Shri Shankaracharya Group of Institutions, Junwani, Bhilai (C.G.), India

2Assistant Professor, Department of Computer Science and Engineering, Shri Shankaracharya Group of Institutions, Junwani, Bhilai (C.G.), India

Abstract: *Clustering problem is an unsupervised learning algorithm. It is a method that partition information objects into matching clusters. The records items inside the identical cluster are quite much like each different and multiple inside the different clusters. Clustering is an unsupervised learning algorithm of hassle that is used to determine the intrinsic grouping in a set of unlabeled statistics. Grouping of gadgets is completed on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity in this kind of way that the items within the same group/cluster share a few similar homes/traits. There is a huge range of algorithms to be had for clustering. This paper provides a comparative analysis of diverse clustering algorithms. In experiments, the effectiveness of algorithms is evaluated through comparing the effects on 4 datasets. Our main aim to show the comparison of the different- different clustering algorithms of WEKA and find out which algorithm will be most suitable for the users.*

Keywords: *clustering, WEKA tool, K-means algorithm, farthest first, etc.*

I. INTRODUCTION

Machine learning studies computer algorithms for learning to do stuff. The goal is to machine learning algorithms that do the learning automatically without human interference. The learning that is being done is always based on some sort of data. So in broad, machine learning is about learning to do better in the future based on what was knowledgeable in the past. There are two types of machine learning – supervised learning and unsupervised learning.

Clustering algorithms are regularly useful in numerous fields like statistics mining, studying theory, sample popularity to discover clusters in a set of facts. clustering is an unsupervised getting to know approach used for grouping factors or facts units in the sort of way that elements in the same group are greater similar (in some way or any other) to every aside from to those in other corporations. Those companies are known as clusters. Clustering is a prime venture of exploratory records mining, and a commonplace method for statistical statistics evaluation, used in lots of fields, along with device studying, sample reputation, image analysis, statistics retrieval, advertising, libraries, insurance, world wide web and bioinformatics. Cluster evaluation become originated in anthropology through motive force and Kroeber in 1932 and introduced to psychology via Zubin in 1938 and Robert Tryon in 1939. Cluster analysis itself is not one precise algorithm, however the well known mission to be solved. It is able to be accomplished by means of various algorithms that range appreciably of their notion of what constitutes a cluster and the way to correctly cluster the factors. Commonly used scheme used to discover similarities among facts elements are inter and intra- cluster distance most of the cluster elements. We will show this with a simple instance:

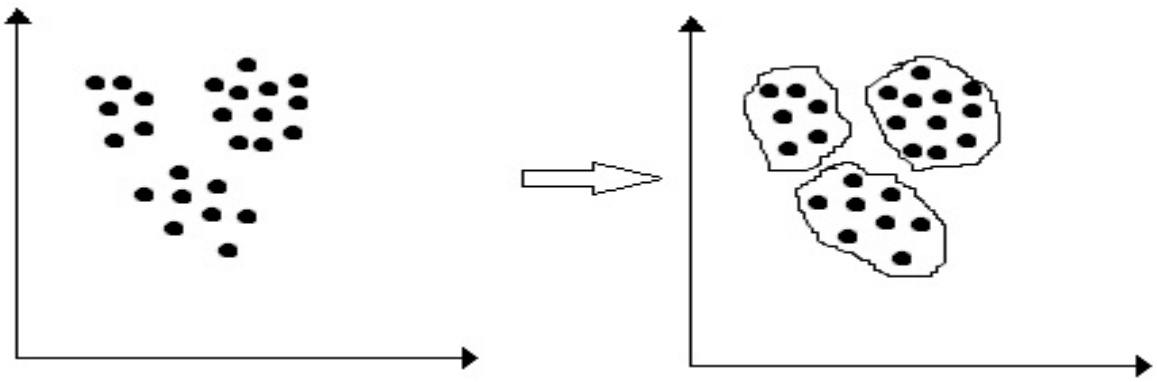

Figure 1 Clustering based on inter and intra distance measure.

Paragraph within the above example, information has been divided into 3 clusters using the similarity criterion "distance": or extra elements belong to the same cluster if they're "nearer" in step with a given distance. For optimizing the clusters, intra-cluster distance should be minimized and inter-cluster distances have to be maximized. This clustering

method is referred to as distance-primarily based clustering. Every other sort of clustering is conceptual clustering where in or extra elements belong to the equal cluster if they are conceptually same or similar.
The ideal clustering algorithm and parameter settings depend on the character facts set and meant use of the outcomes. The diffused variations are often within the usage of the consequences: even as in facts mining, the ensuing organizations are the matter of hobby, in automatic class the ensuing discriminative electricity is of interest. Section 2 of paper presents clustering techniques to be compared. Section 3 gives an overview of WEKA. In section 4 and 5, experimental setup, performance measures and results have been shown. Section 6 concludes the paper.

II. CLUSTERING ALGORITHM

A number of clustering techniques used in data mining tool WEKA have been presented in this section. These are:

2.1 Expectation Maximization

EM algorithm is likewise an essential set of rules of data mining. We used this algorithm whilst we are happy the result of k-method strategies. Expectation– maximization (EM) algorithm is an iterative approach for locating maximum probability or maximum posterior (map) estimates of parameters in statistical fashions, where the version depends on unobserved latent variables. The EM new release alternates among appearing an expectation (E) step, which computes the expectancy of the log likelihood evaluated the usage of the present day estimate for the parameters, and maximization (M) step, which computes parameters maximizing the anticipated log-probability found on the E step. Those parameter-estimates are then used to determine the distribution of the latent variables within the next E step.
The result of the cluster analysis is written to a band named class indices. The values on this band imply the elegance indices, wherein a cost '0' refers to the first cluster; a fee of '1' refers to the second cluster, and many others. The class indices are looked after in keeping with the earlier possibility related to cluster, i.e. a category index of '0' refers back to the cluster with the best probability.

Advantages

1. Offers extraordinarily beneficial end result for the actual world data set.
2. Use this algorithm when you need to perform a cluster analysis of a small scene or vicinity-of interest and aren't satisfied with the consequences acquired from the k-means algorithm.

Disadvantage

1. Algorithm is highly complex in nature

2.2 Density Based Clustering

DBSCAN (for density-based spatial clustering of applications with noise) is a facts clustering algorithm proposed through martin ester, Hans-peter Kriegel, Jorge sander and Xiaowei Xu in 1996 it is a density-based totally clustering set of rules as it finds a number of clusters beginning from the expected density distribution of corresponding nodes. DBSCAN is one of the maximum not unusual clustering algorithms and also most mentioned in medical literature.
OPTICS may be visible as a generalization of DBSCAN to multiple tiers, successfully replacing the parameter with a most search radius. The analysis of DBSCAN in the WEKA is proven within the determine.

Advantage

1. DBSCAN does no longer require you to realize the number of clusters inside the information a priori, as opposed to k-way.
2. DBSCAN can discover arbitrarily shaped clusters. It may even discover clusters completely surrounded by using (however no longer linked to) a extraordinary cluster. Because of the min pts parameter, the so-referred to as single-link effect (distinct clusters being connected with the aid of a skinny line of factors) is decreased.
3. DBSCAN has a perception of noise four. DBSCAN calls for simply two parameters and is generally insensitive to the ordering of the factors in the database. (Best factors sitting on the threshold of two distinctive clusters would possibly change cluster club if the ordering of the factors is modified, and the cluster venture is particular only up to isomorphism.

Disadvantage

1. DBSCAN can most effective bring about a good clustering as accurate as its distance degree is inside the feature region query (p,). The most not unusual distance metric used is the Euclidean distance degree. Mainly for high-dimensional records, this distance metric may be rendered nearly useless due to the so known as "curse of dimensionality", rendering it tough to find the ideal cost for this effect however is present also in some other set of rules primarily based at the Euclidean distance.
2. DBSCAN cannot cluster facts sets nicely with large differences in densities, for the reason that minpts combination can't be chosen as it should be for all clusters then.

2.3 Simple K-mean Clustering

K-mean clustering technique is one of the simplest unsupervised mastering techniques that goal to partition n observations into okay clusters wherein each statement belongs to the cluster with the closest mean cost. To begin with, okay Centroids need to be selected within the beginning. The subsequent step is to take instances or factors belonging to a statistics set and associate them to the nearest facilities. After finding okay new centroids, a new binding has to be performed between the identical facts set factors and the nearest new centre. Manner is repeated until no extra changes are executed. Finally, this algorithm objectives at minimizing intra cluster distance (fee function also called squared blunders function), mechanically inter cluster distance can be maximized.

$$Cost_{Fun} = \sum_{i=1} \sum_{p \in ci} || P - M_i ||^2$$

Where,

M_i – mean of i^{th} cluster,

C_i – i^{th} cluster and

p – Point representing the object.

K-means clustering algorithm is fast, strong, rather green and less complicated to understand. time complexity of the set of rules is O (tknd), wherein n is wide variety of items/ factors inside the records set, ok is number of predefined clusters, d is number of attributes/ measurement of every object, and t is the quantity of iterations till surest clusters aren't received. as it's far a heuristic algorithm, there's no assure that it'll converge to the worldwide ideal and can additionally provide the nearby optima as very last end result depending upon initial cluster centres. Noisy records and outliers are not dealt with.

Advantage

1 With a huge range of variables, k-way may be computationally quicker than hierarchical clustering (if ok is small).

2 K-method may produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

Disadvantage

1 Problem in evaluating satisfactory of the clusters produced (e.g. for one of a kind preliminary partitions or values of okay have an effect on outcome).

2 Constant wide variety of clusters can make it difficult to predict what ok have to be.

3 Does not paintings nicely with non-globular clusters.

2.4 Farthest First Clustering

Farthest first is a heuristic based method of clustering. it's miles a variant of ok way that still chooses centroids and assigns the items in cluster however on the point furthermost from the existing cluster centre lying in the records location. Fast clustering is provided by using this algorithm in maximum of the cases considering the fact that much less reassignment and adjustment is wanted. for each xi = [xi,1, xi,2, …, xi, m] in d that is defined by using m specific d) has been used to denote the frequency rely of attribute value xi, j in the dataset. Then, a scoring characteristic has been designed for evaluating each point, which is defined as:

$$Score\ (X_i) = \sum^{m} f(X_{i,j} \mid D)$$

In the farthest-factor heuristic, the point with highest score is selected because the first point and closing points are decided on in the equal manner as that of fundamental farthest-point heuristic. Deciding on the first point in keeping with above described scoring function can be fulfilled in o (n) time by deploying the subsequent method (with scans over the dataset):

(1) In the first test over the dataset, m hash tables are constructed as basic statistics structures to store the statistics on characteristic values and their frequencies in which m is variety of attributes.

(2) Inside the second scan over the dataset, with using hashing technique, in o (1) predicted time, the frequency be counted of an attribute value in corresponding hash table may be determined.

Consequently, the statistics factor with largest score might be detected in o (n) time. Time complexity of the basic set of rules is o (nk), where n is quantity of items inside the dataset and okay is variety of desired clusters. In basic of clustering, first point is chosen randomly. Farthest-point heuristic based technique is appropriate for big-scale records mining packages.

Advantage

Farthest-point heuristic based method has the time complexity o (nk), in which n is variety of objects within the dataset and okay is quantity of favoured clusters. Farthest-factor heuristic primarily based method is rapid and suitable for large-scale data mining applications.

III. WEKA

WEKA (Waikato Environment for Knowledge Evaluation) is an open source, platform impartial and smooth to use statistics mining device issued beneath gnu general public license. It comes with Graphical User Interface (GUI) and incorporates series of information Pre-processing and Modelling strategies. Gear for facts pre-processing, classification, regression, clustering, association policies and visualization in addition to appropriate for new device gaining knowledge

of schemes are furnished inside the bundle. It's far portable due to the fact that it's miles absolutely implemented inside the java programming language and for that reason runs on nearly any contemporary computing platform.

User interfaces

WEKA's fundamental person interface is the explorer, but essentially the same functionality may be accessed through the factor-primarily based information glide as well as the command line interface (CLI). There may be additionally the experimenter, which allows the systematic evaluation of the predictive performance of WEKA's device getting to know algorithms on a group of datasets. The explorer interface capabilities numerous panels imparting get admission to the main additives of the workbench:

- The pre-process panel has facilities for importing statistics from a database, a csv or an arff document, and many others and for pre-processing this information using a so-known as filtering algorithm. Those filters may be used to convert the facts from numeric to discrete, to take away lacking instances, to correctly pick out lacking values and converting csv record to arff and vice versa.
- The classify panel allows the user to apply class and regression algorithms to the ensuing dataset, to estimate the accuracy of the resulting predictive version, and to visualize errors. there are various form of type algorithms like rule primarily based, decision tree, naïve Bayesian, lazy, mi, misc etc. this paper make use of selection tree category algorithms.
- The partner panel attempts to pick out all critical interrelationships among attributes within the facts with the assist of association freshmen like Apriori, filtered associate, predictive apriori etc.
- The cluster panel offers get admission to the clustering strategies in WEKA, e.g., the simple k-approach, , CLOPE set of rules to provide one of a kind type of clustering's for exclusive conditions and usage in their effects.
- The pick attributes panel offers algorithms for identifying the most predictive attributes in a dataset.
- The visualize panel indicates a scatter plot matrix, wherein person scatter plots can be decided on and enlarged, and analyzed in addition using diverse selection operators.

Extension packages

In version 3.7.2 of WEKA, a package manager was added to allow the easier installation of extension packages. Much functionality has come in WEKA through continuous extension and updates to make it more sophisticated.

IV. METHODOLOGY & PERFORMANCE MEASURES

Clustering techniques discussed in section 3 have been compared with the help of WEKA. **Performance measure** used to determine accuracy of clustered data is **class to cluster evaluation**. A little about some important terms which are used in this measures is presented. These are:-

- True Cluster (TC) – Total number of elements belonging to clusters that were correctly predicted. These elements are verified using their classes i.e. TC= TC1 + TC2 + … TCn. Here n is the number of classes in the dataset and TCi is the number of elements of class Ci which belongs to correct/right cluster.
- N – Total number of instances which are clustered.

Accuracy: It determines the proportion of the total number of instances clustered to the instances which are correctly clustered.

$$\text{Accuracy} = TC/N$$

V. EXPERIMENTAL RESULTS

A comparative analysis of diverse clustering algorithms has been made using six datasets taken from the keel (a software tool to evaluate evolutionary algorithms in records mining problems) and UCI machine getting to know repository. All of the datasets are summarized in table 1.

Table -1: Datasets used in Experiments

DATASET	INTANCES	ATTRIBUTES	CLASSES
TIC TAC TOE	958	10	2
BREAST CANCER	277	10	2
CAR	1728	7	4
MASROOM	5644	23	2

Results are observed using two measures; accuracy and time, explained in section using all the datasets mentioned in Table 1. Results have been shown in the Table 2, 3, 4, and 5.

Table – 2: Comparison of Various Clustering Algorithms for Tic Tac Toe Dataset.

Clustering Method	Accuracy	Time Taken (in sec.)
Expectation Maximization	36.75	19.19
Make Density Based Clustering	54.38	0.04
Hierarchical Clustering	65.14	6.52
Simple K-means Clustering	50.53	0.02
Farthest First Clustering	55.75	0.01

Table -3: Comparison of Various Clustering Algorithms for Breast Cancer Dataset.

Clustering Method	Accuracy	Time Taken
Expectation Maximization	68.71	1.91
Make Density Based Clustering	73.42	0.01
Hierarchical Clustering	70.62	0.34
Simple K-means Clustering	74.47	0.01
Farthest First Clustering	65.73	0.01

Table -4: Comparison of Various Clustering Algorithms for Car Dataset.

Clustering Method	Accuracy	Time Taken
Expectation Maximization	70.02	3.62
Make Density Based Clustering	67.18	0.02
Hierarchical Clustering	69.96	83.45
Simple K-means Clustering	67.18	0.04
Farthest First Clustering	46.58	0.01

Table -5: Comparison of Various Clustering Algorithms for Mushroom Dataset.

Clustering Method	Accuracy	Time Taken
Expectation Maximization	42.54	983.55
Make Density Based Clustering	58.32	0.34
Hierarchical Clustering	66.16	0.70
Simple K-means Clustering	62.38	0.24
Farthest First Clustering	60.61	0.06

Within the evaluation, unique measures have been used for comparing numerous clustering algorithms. From the effects received inside the tables 2, 3, 4, and 5. It can be seen that farthest first performs fine among all in maximum of cases. Clustering accuracy in farthest first is maximum and time taken in clustering is minimal. Expectation maximization clustering has validated worst in all the cases. Its clustering accuracy is minimum in addition to time taken is maximum. Relaxation of the models lies in between the exceptional and worst ones.

4. CONCLUSION

Within the current few years data mining techniques covers every location in our lifestyles. We are the usage of information mining strategies in especially in the clinical, banking, insurances, training etc. before start operating in the with the facts mining fashions, it's far very vital to understanding of available algorithms. The principal purpose of this paper to offer an in depth advent of WEKA clustering algorithm. WEKA is the records mining tools. It's far the best device for classify the facts diverse sorts. It is the primary model for offer the graphical consumer interface of the consumer. It is providing the past project data for analysis. Comparative analysis of diverse clustering algorithms has been made. The results were validated the use of four datasets taken from UCI and keel repository and observed that datasets are successfully clustered with a quite suitable accuracy. Few of the clustering techniques have better accuracy, others take less time, and many others have a trade-off between accuracy and time taken. Suitable methods can be used in keeping with their utilization.

ACKNOWLWDGEMENT

I want to thank my guide prof. Yamini Chouhan and the management of Shri Shankaracharya Group of Institute and Technical Campus, Junwani, Bhilai, India; for their inspiring encouragement and support towards the completion of this research.

REFERENCES

[1] Narendra Sharma, Aman Bajpai, Mr. Ratnesh Litoriya “Comparison the various clustering algorithms of WEKA tools” International Journal of Emerging Technology and Advanced Engineering Volume 2, Issue 5, May 2012

[2] Priyanka Sharma, “Comparative Analysis of Various Clustering Algorithms Using WEKA” , International Research Journal of Engineering and Technology (IRJET) Volume: 02 Issue: 04 | July-2

[3] S.Revathi, Dr.T.Nalini, “Performance Comparison of Various Clustering Algorithm” International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 2, February 2013

[4] ZHEXUE HUANG, “Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values” Kluwer Academic Publishers. Manufactured in the Netherlands 1998

[5] Jinxin Gao, David B. Hitchcock, “James-Stein Shrinkage to Improve K-means Cluster Analysis”

[6] Istvan Jonyer, Diane J. Cook, Lawrence B. Holder, “Graph-Based Hierarchical Conceptual Clustering” Journal of Machine Learning Research 2 (2001)

[7] Sean Borman, “The Expectation Maximization Algorithm A short tutorial” July 18 2004

[8] Garima Sehgal1, Dr. Kanwal Garg, “IMPROVED EXPECTATION MAXIMIZATION CLUSTERING ALGORITHM” International Journal Of Engineering And Computer Science Volume 3 Issue 5 may, 2014

[9] Garima Sehgal#1 Dr. Kanwal Garg, “Comparison of Various Clustering Algorithms” International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014,

[10] Sanjoy Dasgupta, “Performance guarantees for hierarchical clustering” Preprint submitted to Elsevier Science 24 July 2010

[11] Nidhi Suthar, Prof. Indr jeet Rajput, Prof. Vinit Kumar Gupta “A Technical Survey on DBSCAN Clustering Algorithm,” international journal of scientific & Engineering Research, volume 4, issue 5, May 2013.

[12] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, jorg Sandar “OPTICS: Ordering Points to Identify the Clustering Structure,” Proc. ACM SIGMOD’99 Int. Conf. on Management of data, Philadelphia PA, 1999.

[13] Martin Ester, Hans-Peter Kriegel, Jorg sander, Xiaowei Xu "A Density Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," international conference on knowledge discovering and data mining.

[14] Huang Darong, Wang Peng, "Grid-based DBSCAN Algorithm with Referential Parameters" International Conference on Applied Physics and Industrial Engineering, 2012

[15] Yuni Xia, Bowei Xi, "Conceptual Clustering Categorical Data with Uncertainty" 19th IEEE International Conference on Tools with Artificial Intelligence, 2007

[16] A. P. Dempster, N. M. Laird; D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm" Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, Apr 6 2007

[17] E. B. Fowlkes and C. L. Mallows, "A Method for Comparing Two Hierarchical Clustering" Journal of the American Statistical Association, Vol. 78, 20/05/2010

[18] Deepshree A. Vadeyar1,Yogish H.K, "Farthest First Clustering in Links Reorganization" International Journal of Web & Semantic Technology (IJWEST) Vol.5, No.3, July 2014

[19] Mamta Mor1, Poonam Gupta2, Priyanka Sharma, "A Genetic Algorithm Approach for Clustering" International Journal of Engineering And Computer Science Volume 3 Issue 6 June, 2014

[20] Zengyou He, "Farthest-Point Heuristic based Initialization Methods for K-Modes Clustering"

[21] Amey K. Redker, prof. S. R. Todmal, "A Survey on DBSCAN Algorithm to Detect Cluster with Varied Density," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) volume 5, issue7, july 2016.

