ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 7 | Issue: 8 | August 2021|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2021: 8.047 || ISI Value: 1.188

A COMPARATIVE ANALYSIS OF K-MEANS AND HIERARCHICAL CLUSTERING

Aastha Gupta[1], Himanshu Sharma[2], Anas Akhtar[3]

[1,2,3] *Jagan Institute of Management Studies, sec-5, Rohini*

Article DOI: https://doi.org/10.36713/epra8308
DOI No: 10.36713/epra8308

ABSTRACT

Clustering is the process of arranging comparable data elements into groups. One of the most frequent data mining analytical techniques is clustering analysis; the clustering algorithm's strategy has a direct influence on the clustering results. This study examines the many types of algorithms, such as k-means clustering algorithms, and compares and contrasts their advantages and disadvantages. This paper also highlights concerns with clustering algorithms, such as time complexity and accuracy, in order to give better outcomes in a variety of environments. The outcomes are described in terms of big datasets. The focus of this study is on clustering algorithms with the WEKA data mining tool. Clustering is the process of dividing a big data set into small groups or clusters. Clustering is an unsupervised approach that may be used to analyze big datasets with many characteristics. It's a data-modeling technique that provides a clear image of your data. Two clustering methods, k-means and hierarchical clustering, are explained in this survey and their analysis using WEKA tool on different data sets.

KEYWORDS: ***data clustering, weka , k-means, hierarchical clustering***

I. INTRODUCTION

Clustering is a vital part of data mining, and it's also one of the hottest topics of science in recent times. It is a technology that examines the logical or physical relationships between data and divides the data set into many clusters, each of which is made up of similar data sets in nature.

Data clustering is a process in which we group together entities with similar characteristics. Clustering quality depending on the similarity metric and how it's implemented. The clustering's main aim is to find a collection of patterns, points, and connections or objects from a natural grouping. One of the most remarkable data mining technique is clustering. Based on some rules, data may be classified into several classes or clusters, resulting in great similarity among data sets of the same class and substantial differences among data objects of other classes. [1]

Clustering is a method for logically categorizing raw data and looking for hidden patterns in large datasets. It's the act of grouping data into fragmented clusters so that data in one cluster matches data in another, while data in other clusters varies. Clustering is a common data analysis approach for identifying homogenous groups of objects based on attribute values. Data Clustering has many different real life applications such as image segmentation, data analysis, machine learning, search engines, document retrieval, object recognition and evaluation, computational, economics, libraries, insurances studies.

Clustering algorithms are effective meta-learning tools for assessing the information generated by modern applications. Clustering methods are widely employed in a variety of applications. Data organization and categorization, as well as data modelling as well as data compression. When selecting a clustering algorithm, think about whether it can scale to your dataset. Machine learning datasets can contain millions of instances, but not all clustering algorithms scale well. The similarity of all pairs of examples is computed by several clustering algorithms.

Clustering approaches are used to classify groups of related data in multivariate data sets. There are a variety of clustering methods including:

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 7 | Issue: 8 | August 2021|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2021: 8.047 || ISI Value: 1.188

- Partitioning methods
- Fuzzy clustering
- Hierarchical clustering
- Density-based clustering
- Model-based clustering

II. LITERATURE REVIEW

1. Manish Verma, Mauly Srivastava, Neha …" A Comparative Study of Various Clustering Algorithms in Data Mining" [5]

The author made a comparison between different clustering techniques. The aim was to measure the algorithm which gives the best performance. It was observed that K-means is faster than all the algorithms that are mentioned in this paper. K-means and EM gives the best results than hierarchical clustering when working on huge data set.

2. U. Kaymak and M. Setnes, "Extended fuzzy clustering algorithms" [6]

The author uses fuzzy clustering algorithm to divide dataset into clusters. Some of the issues using fussy algorithm were discussed by the author such as number and shape of clusters, division of data patterns, choosing the number of clusters in the data. Enhanced version of fussy means were given and their properties were illustrated. Examples were used to show that the enhanced algorithms does not require any additional input from the user and can determine partition of data on its own.

3. Karthikeyan B., Dipu Jo George, G. Manikandan, Tony Thomas "A comparative study on k-means clustering and agglomerative hierarchical clustering," [7]

The authors have done a comparative study to determine the best-suited algorithm among K-Means and Agglomerative Hierarchical Clustering. It was concluded that k-means can be best used for larger datasets with minimal runtime and memory change rate. It is also concluded that the agglomeration hierarchical clustering technique is best suited for smaller data sets because of the minimum overall memory consumption.

4. S. H. Sastry, P. Babu and M. S. Prasada, "Analysis & Prediction of Sales Data in SA P-ERP System" using Clustering Algorithms", [8]

The authors of this paper used grouping procedures for recognizing contrast in item deals and furthermore to recognize and think about deals throughout a specific time. The interest for steel items is repeating furthermore, relies upon numerous variables like client profile, value, limits and expense issues. Creators have investigated deals information with bunching calculations like K-Means and EM (assumption augmentation) that uncovered many fascinating examples helpful for improving deals income and accomplishing higher deals volumes. K-Means and EM (segment Procedures) calculations are more qualified to assess deals information in correlation with thickness based Procedures.

5. Soumi Ghosh, S. K. Dubey, "Comparative Analysis of K-Means….." [9]

The paper includes comparison of two clustering techniques, centroid-based K-Means and representative object-based Fuzzy C-Means clustering techniques. This analysis is based on a performance evaluation with these algorithms about how efficient outputs are generated. The results of this comparative research depicts that efficiency of FCM is somewhat closer to K-means. However, computation time is still longer than K-means since the fuzzy measure calculations are involved.

6. M.Venkat Reddy, M. Vivekananda, RUVN Satish. [10]

The researchers have discovered an efficient clustering technique by comparing Divisive and Agglomerative Hierarchical Clustering with K-means. The outcome of paper was that Agglomerative clustering along with k-means is the practical choice to achieve a high degree of accuracy. Divisive clustering with k-means also functions efficiently where each cluster is fixed i.e. where the initial centroids are taken in a fixed number for each cluster rather than by random selection.

7.. N. Sharma ."Comparison the various clustering algorithms of weka tools".[11]

The authors have compared and contrasted different clustering algorithms. Weka Tool is used to implement all of the proposed algorithms. The purpose of their research is to determine which algorithm is more appropriate and efficient. DBSCAN, EM, Farthest First, OPTICS, and the K-Means algorithms are among these algorithms. They show the benefits and drawbacks of each algorithm in this study. They have demonstrated the benefits and drawbacks of each method in this paper, however based on their study, they discovered that the k-means clustering algorithm is the simplest of the algorithms and fastest algorithm to be used with large datasets.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 7 | Issue: 8 | August 2021|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2021: 8.047 || ISI Value: 1.188

III. CLUSTERING PROCESS

The analytical processes required in cluster analysis have been established in the literature based on the basic paradigm on Knowledge Discovery in databases. Figure 1 depicts the steps involved in the clustering process.[3]

1. Feature selection

The stage is about choosing characteristics for cluster analysis. Because the class labels aren't predefined in cluster analysis, there's a good chance you'll pick features that are irrelevant or inconsequential. Additionally, removing non-essential information improves clustering results. The process of determining the most effective subset of the original characteristics to employ in clustering is known as feature selection. The application of one or more modifications of the input features to create new salient characteristics is known as feature extraction. To get an adequate collection of characteristics to employ in clustering, one or both of these strategies can be applied.

2. Clustering algorithm

The choice of a clustering algorithm influences the clusters obtained from the data . The results obtained from clustering algorithms are based on some assumptions which depends on the properties of the data set (geometry and density distribution) and input parameter values since the class labels are not specified. A good clustering algorithm can recognize clusters regardless of their structure.

3. Cluster validation

Cluster validation of the clusters is an assessment of the clusters generated. Clusters are checked to determine a satisfactory quality of the created clusters and to achieve the desired clusters. External clusters can all be used to test clusters with internal indices and relative indices. The clusters generated by the algorithm are assessed in this stage. Visualizing the clusters is a useful way to rapidly double-check the cluster results. [4]

4. Result Analysis

The clusters produced from the initial set of data are analyzed to gain a better understanding of them and to guarantee that the attributes of the clusters are obtained. Integration of expert evaluations with additional experimental findings and analysis might also help to broaden the interpretation.

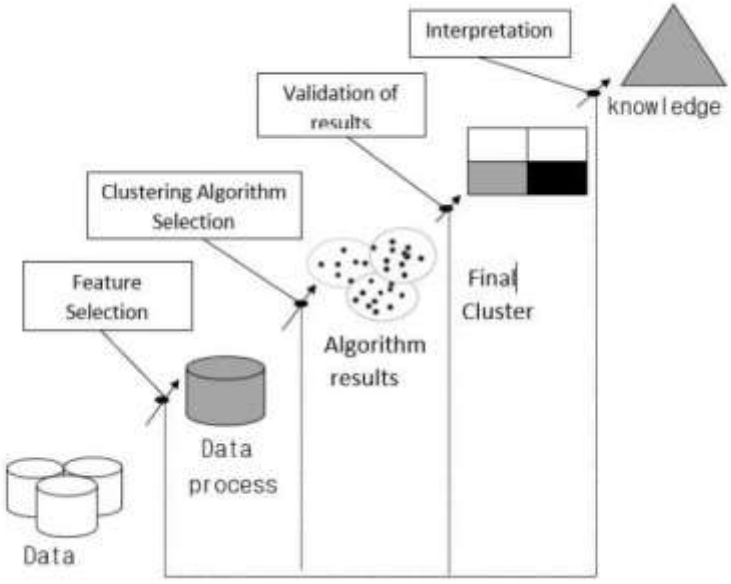


Fig 1 . Clustering Process

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 7 | Issue: 8 | August 2021|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2021: 8.047 || ISI Value: 1.188

IV. CLUSTERING BENCHMARKING CRITERION

The comparative strengths and limitations of each algorithm in relation to the three-dimensional [3-D] characteristics of large data should be analyzed by particular criteria for the evaluation of large-data clustering methods including Volume, Velocity, and Variety.

The efficiency to manage a large amount of data is called volume of a clustering process. The following criteria are taken into account while choosing a good clustering algorithm for the Volume property:

i) the dataset's size,
ii) dealing with high dimensionality, and
iii) managing the noisy data.

The capability to handle various sorts of data is referred to as variety of clustering process. The following criteria are taken into account while choosing a good clustering algorithm for the Variety property:

i) the dataset type;
ii) shape of clusters

The speed of an algorithm over massive data is referred to as velocity of clustering process. The different criteria are taken into account while choosing a good clustering procedure for the Velocity property:

i) the algorithm's complexity;
ii) the algorithm's run-time performance

V. COMPARITIVE ANALYSIS

V.I. K-Means

The K means clustering algorithm is commonly used. This technique will be useful in extracting meaningful information from a large database using a cluster. The K-means clustering algorithm is a well-known data clustering technique. It is used in a variety of applications, including information retrieval and computer vision. K-means clustering divides n data points into k clusters, allowing for the grouping of comparable data points. It's an iterative strategy for assigning each point to the cluster with the closest centroid. The centroid of these clusters is then calculated again by taking the average.

Advantages

- Simple: - Easy to understand and to implement.
- Efficient: Time complexity is O(t.k.n) very efficient to work with huge data sets
- Requires an input from user.

Disadvantages

- K-Means may be computationally faster only if value of K is small.
- Can only be used if the mean is known.
- Not suitable for high dimensional data
- Sensitive to noise/outliers [12]

V.II. Hierarchical clustering

A hierarchical method creates a hierarchical representation of a set of data items. Dendrograms are made using the Tree of Clusters. Sibling clusters split the points covered by their shared parent, whereas child clusters exist in every cluster node. A typical clustering approach that can be helpful for a range of data mining tasks is hierarchical algorithms. A hierarchical clustering technique creates a succession of clusterings in which each grouping gets nestled into the clustering behind it.

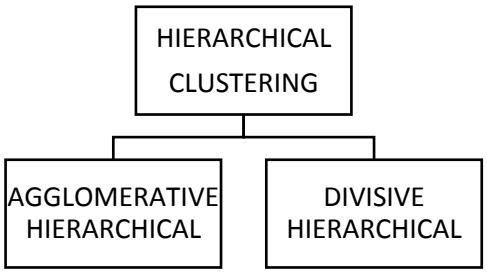
Advantages

- Applicable to all attribute types.
- Easy at handling similarity data.
- Small groups are formed, making analysis and comprehension simpler.
- The number of clusters are not pre-defined, so the user has the ability to dynamically select clusters.
- Concept wise simple.

Disadvantages

- Clustering Cluster merging/splitting is a permanent process.
- It is impossible to correct erroneous judgments afterwards.
- Divisive techniques can be time-consuming to compute.
- Methods aren't always (necessarily) scalable when dealing with huge datasets.
- A termination/readout condition is required.

Hierarchical clustering can be divided into two sub categories:



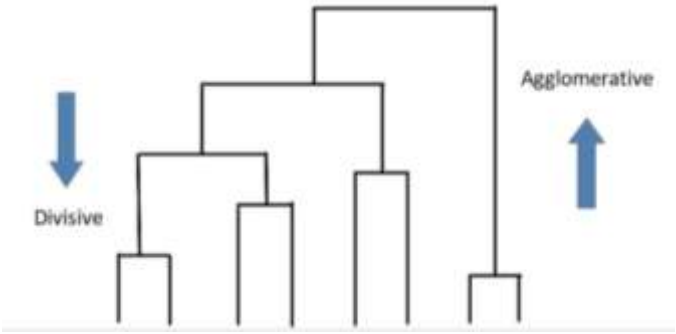
ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 7 | Issue: 8 | August 2021|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2021: 8.047 || ISI Value: 1.188

I. Agglomerative Hierarchical clustering

The bottom-up approach is often referred to as the agglomerative approach, since it begins with each object that forms a separate group. It continues to merge the nearby objects or groups. It keeps going until all the groups are combined to one or until the condition of termination is maintained. The aim of agglomerative clustering technique is to group together objects with similar characteristics. [14]

II. Divisive Hierarchical clustering

The divisive clustering method, on the other hand, works from the top down, starting with a single cluster at the top and dividing it down to the bottom. It usually starts in the same cluster with all of the objects. Then, through the application of the K-means clustering, a cluster is divided into smaller clusters. It is down until the termination condition carries every object in one cluster. [13]



VI. WEKA TOOL

Weka is freely available on the Internet and comes with a new data mining document that describes and thoroughly explains all of the techniques that are included. Weka class libraries-based applications may operate on any computer with a Web browser, allowing users to apply machine learning algorithms to their own data, independent of computer's platform. We used the Weka tool version 3.8.5 in this work to examine the accuracy and speed of simple K-means and Hierarchical clustering algorithms on pre-given datasets.

VII. EXPERIMENT

Various datasets with known clustering are available in the UCI collection of machine learning databases for testing the accuracy and efficiency of simple k-means and hierarchical clustering algorithms. The Diabetes datasets and Hypothyroid datasets, as well as a brief explanation of datasets utilized in experiment evaluation, are used in this study. [11]

Table 1 lists some of the features of the test datasets – number of attributes and number of instances formed in the given dataset.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 7 | Issue: 8 | August 2021|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2021: 8.047 || ISI Value: 1.188

Table 1. Description of Data Sets

Datasets	No. of Attributes	No of Instances
Diabetes	09	768
Hypothyroid	30	3772

Table 2.Clustering Results for Data Sets.

	k-means running time(sec)	Hierarchical clustering running time (sec)	time k-means Accuracy %	Hierarchical clustering Accuracy %
Diabeties	0.06	2.14	51.692	65.104
Hypothyroid	0.16	5.74	69.64	93.24

Table 2 shows the clustering findings for cluster k=3.

Fig 2 Shows Running Time when Both Algorithms are applied on the Same Datasets.

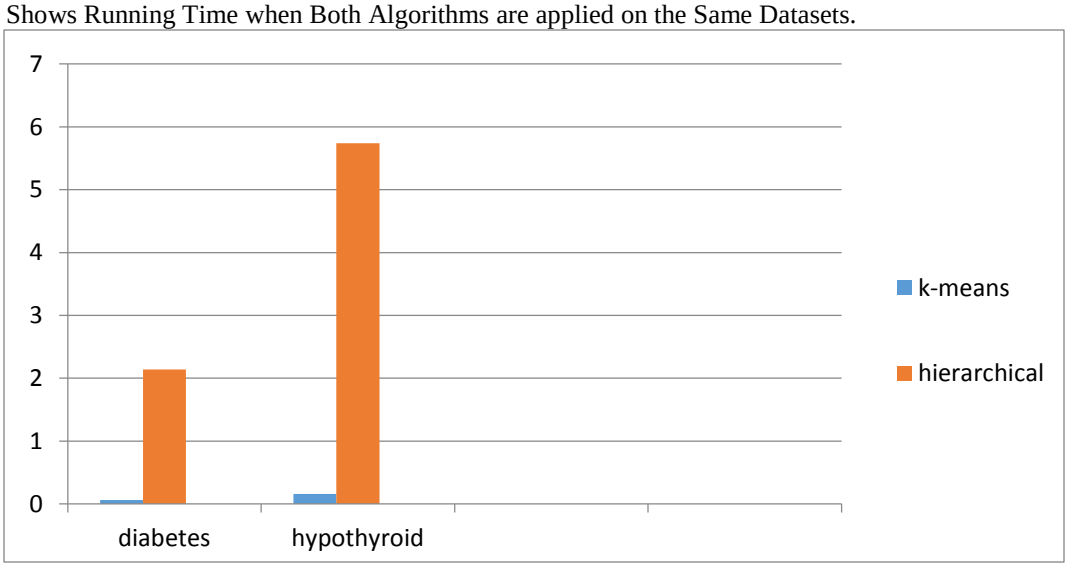


Figure 2. Running time v/s Datasets

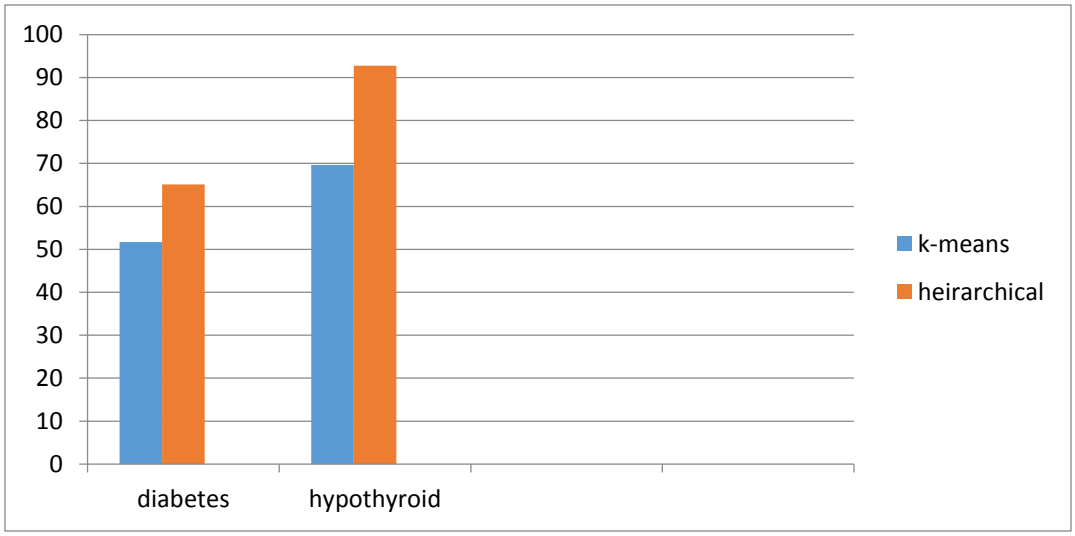


Figure 3. Accuracy v/s Datasets

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 7 | Issue: 8 | August 2021|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2021: 8.047 || ISI Value: 1.188

VIII. CONCLUSION

The K-mean method performs in clustering huge data sets, and its performance improves as the number of clusters grows. For categorical data, a hierarchical algorithm was employed, and according to its complexity, a new approach for giving rank values to each categorical attribute using K- means was applied, in which categorical data is first transformed to numeric by assigning rank values to each categorical attribute. The K-mean algorithm performs better than the Hierarchical Clustering Algorithm. The RMSE lowers as the number of clusters rises, and the performance of the K-Means method improves as the RMSE drops. When clustering certain (noisy) data, all of the methods contain some ambiguity. When clustered, all of the methods exhibit some uncertainty in some (noisy) data. When a large dataset is used, the quality of all algorithms improves dramatically. The K-Means algorithm is extremely sensitive to dataset noise. This noise makes it difficult for the algorithm to cluster data into appropriate clusters, and thus has an impact on the method's outcome. When working with large datasets, the K-Means system results conventional clustering algorithms while still producing high-quality clusters.

IX. REFERENCES

1. *D. Karaboga and C. Ozturk, "A novel clustering approach: Artificial Bee Colony (ABC) algorithm",Applied Soft Computing, vol. 11, no. 1, **(2011)**, pp. 652-657.*
2. *J. Senthilnath, S. N. Omkar and V. Mani, "Clustering using firefly algorithm: performance study",Swarm and Evolutionary Computation, vol. 1, no. 3, **(2011)**, pp. 164-171.*
3. *M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques, " J. Intell. Inf. Syst., vol. 17, no. 2–3, pp. 107–145, 2001.*
4. *K. Wang, B. Wang, and L. Peng, "CVAP: validation for cluster analyses," Data Sci. J., vol. 8, pp. 88–93, 2009.*
5. *M. Verma, M. Srivastava, N. Chack, A. K. Diswar, and N. Gupta, "A Comparative Study of Various Clustering Algorithms in Data Mining," Int. J. Eng. Res. Appl.*
6. *U. Kaymak and M. Setnes, "Extended fuzzy clustering algorithms", ERIM Report Series Reference No.ERS-2001-51-LIS, **(2000)**.*
7. *B. Karthikeyan, D. J. George, G. Manikandan, and T. Thomas, "A comparative study on k-means clustering and agglomerative hierarchical clustering," Int. J. Emerg. Trends Eng. Res., vol. 8, no. 5*
8. *S. H. Sastry, P. Babu and M. S. Prasada, "Analysis & Prediction of Sales Data in SA P-ERP System using Clustering Algorithms", arXiv preprint arXiv:1312.2678, **(2013)**.*
9. *Soumi Ghosh, Sanjay Kumar Dubey, Comparative Analysis ofK-Means and Fuzzy C-Means Algorithms, International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013.*
10. *M. V. Reddy, M. Vivekananda, and R. U. V. N. Satish, "Divisive Hierarchical Clustering with K-means and Agglomerative Divisive Hierarchical Clustering with K-means and Agglomerative Hierarchical Clustering,"*
11. *N. Sharma, A. Bajpai and R. Litoruya, "Comparison the various clustering algorithms of weka tools" International Jornal of Emerging technology and Advanced Engineering, vol. 2, no. 5, (2012) May.*
12. *Amit Saxena1 , Mukesh Prasad2 , Akshansh Gupta3 , Neha Bharill4 , Om Prakash Patel4 , Aruna Tiwari4, A Review of Clustering Techniques and Developments*
13. *K. Wang, B. Wang, and L. Peng, "CVAP: validation for cluster analyses," Data Sci. J., vol. 8, pp. 88–93, 2009*
14. *Performance of selected agglomerative hierarchical clustering methods nusa erman1 , ales korosec2 , jana suklan3*