ISSN: 2349 - 4891
International
Journal of Recent Research and Applied Studies
(Multidisciplinary Open Access Refereed e-Journal)

A Review on Clustering Algorithms

Dr. J. Charles Selvaraj[1] & Dr. V. Arul Kumar[2]

[1]*Assistant Professor, A. A. Govt. Arts & Science College, Musiri, Tamilnadu, India.*
[2]*Assistant Professor, Thanthai Hans Roever College (Autonomous), Perambalur, Tamilnadu, India.*

Received 20th June 2016, Accepted 5th August 2016

ABSTRACT

Data mining is an automatic data analysis method to uncover the previously undetected relationship, meaningful patterns and the rules among data items. This technique generally performs the analysis over the large volume of data stored in database, data warehouse and various information repositories. Data mining is the part of the Knowledge Discovery process. Clustering is an important technique to classify the unlabeled data. The data in the same cluster are more similar to each other and then to those in other cluster. In this clustering technique grouping of data became very complex task due to the availability of more number of features and the clustering accuracy gets degraded. In the field of data mining feature selection plays a vital role to select the most relevant features. This research article gives overview on clustering algorithms and it is very useful to the budding researchers in the respective field.

Keywords: Data Mining, Clustering, Feature Selection, Various statistical measures, Clustering Algorithm.

I. INTRODUCTION

Data mining is an automatic data analysis method to uncover the previously undetected relationship, meaningful patterns and the rules among data items. This technique generally performs the analysis over the large volume of data stored in database, data warehouse and various information repositories. Data mining is the part of the Knowledge Discovery process. Knowledge discovery in data bases frequently abbreviated as KDD. Data mining plays an important role in the KDD process and it consist of several steps: such as Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern Evaluation and Knowledge Presentation. The most important data mining techniques are classification, clustering, association rule mining, pattern evaluation and regression [1]. In the past decades huge volume of data are generated and stored in digital format. The presence of a large number of features has become a very complex task to perform the clustering process. Feature Selection is an important preprocessing technique in data mining. The feature selection is the process of selecting the relevant features from the original dataset. The objectives of the feature selection technique are to reduce the number of features and to improve the clustering accuracy [2].

Correspondence
Dr.V.Arul Kumar
E-mail: arulkumarvenugopal@gmail.com, Ph. +9189036 80533

II. FEATURE SELECTION

Feature Selection is a process of finding an optimal or suboptimal subset of x features from the original X features. It requires a large search space to obtain the optimal feature subset. The optimal feature subset is measured by evaluation criteria [3].

A. General Feature Selection Procedure

The feature selection procedure includes four important key steps; subset generation, subset evaluation, stopping criterion and result validation [4] which is shown in the following figure 1.

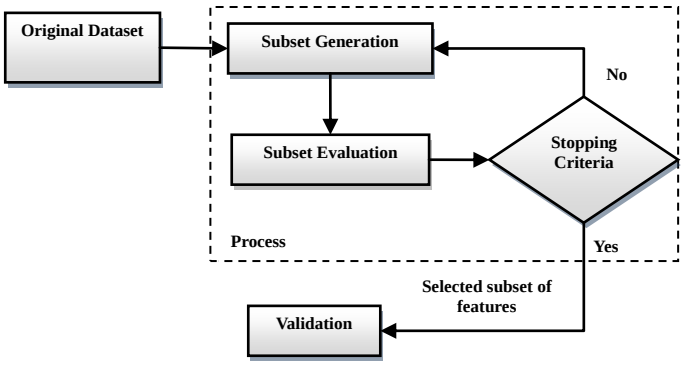

Figure 1. General Feature Selection Procedure

- *Subset Generation*

It is a search process that generates the candidate feature subset using certain search strategy. The process has two basic issues. They are, search direction and search strategy. Firstly, a starting point must be selected which in turn influences the search direction. The search directions are divided into forward search, backward search and bi-directional search [Liu, 1998]. The search process starts with an empty set and adds the features progressively one by one (forward

search) or starts with full sets and removes the features one by one (backward search) or starts with both ends and adds and removes the features simultaneously(bi-directional search). Secondly, a search strategy must be decided. The search strategies are categorized into complete search, sequential search and random search [4].

- *Subset Evaluation*

In subset evaluation, the evaluation criterion is used to evaluate each newly generated subset. The evaluation criterion is used to determine the goodness of the subset (i.e., an optimal subset selected using one criterion may not be optimal according to another criterion). The evaluation criteria are divided into Independent, Dependent and Hybrid criteria [4].

- *Stopping Criterion*

It is used to stop the feature selection process. The feature selection process may stop under one of the following criteria [4].

- A predefined number of features is selected,
- A predefined number of iterations is reached,
- In case, addition (or deletion) of a feature fails to produce a better subset,
- An optimal subset according to the evaluation criterion is obtained.

- *Validation*

The validation process is used to measure the resultant subset using the prior knowledge about the data. In some applications, the relevant features are known beforehand, a comparison is done between the known set of features with the selected features [4].

B. Feature Selection Approaches

The feature selection approaches are broadly classified into three types. They are, Filter Approach, Wrapper Approach and Hybrid Approach [5]

- *Filter Approach*

The filter approach shown in Fig 2 is used to find the relevance of features by looking at the intrinsic properties of the data. A feature is selected independently from the learning method that uses the selected features. In this process, irrelevant features are filtered out. This approach is computed easily and very efficiently [6].

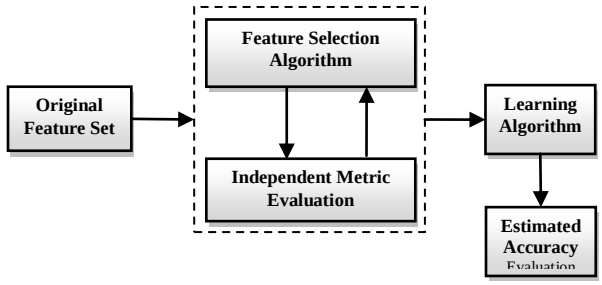

Figure 2. Filter Approach in Feature Selection

- *Wrapper Approach*

The Wrapper approach shown in Fig 3. The features are selected with the use of learning algorithm. The accuracy of the features selected by the filter approach is less, whereas in wrapper approach the accuracy is high. The computational speed is higher in filter approach compared to wrapper approach [7].

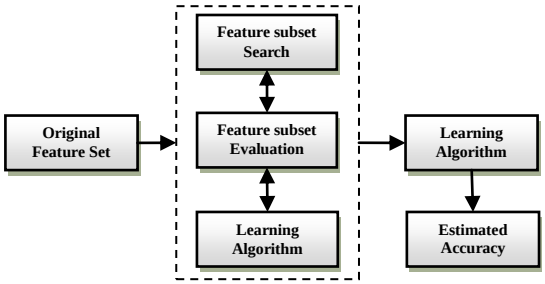

Figure 3. Wrapper Approach in Feature Selection

- *Hybrid Approach*

The Hybrid approach is developed by combining the above filter approach and wrapper approach to handle larger datasets. In this approach the feature set is evaluated using both independent measure and a data mining algorithm. The independent measure is used to choose the best subset for a given cardinality and the data mining algorithm selects the finest subset among the best subsets across diverse cardinalities [8]. Fig 4 shows the Hybrid approach.

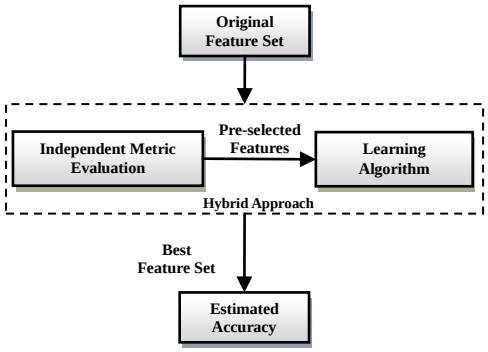

Figure 4. Hybrid Approach in Feature Selection

III. CLUSTERING

Clustering is one of the most widely used techniques to identify a meaningful group among the data points without knowing its structure [9]. The following fig 5. Shows the data flow representation in clustering technique

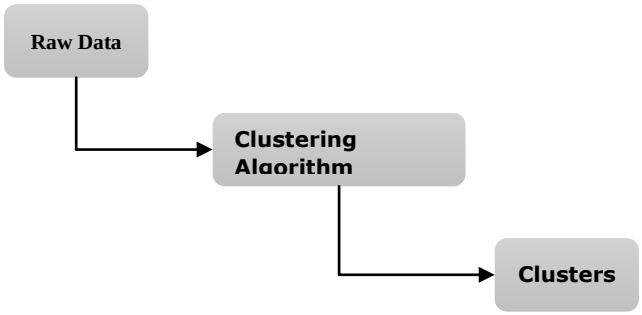

Figure 5. Data Flow Representation in Clustering Technique

IV. EXISTING CLUSTERING ALGORITHM

In this section various clustering algorithms are analysed for the improvement of the clustering accuracy which were developed in the past decades.

A. K MODES ALGORITHM

The K modes algorithm [17] is a used for partition the data into various clusters. In this algorithm mode value is considered to and groups the clusters using the frequency based methods. The following steps shows the methodology of this algorithm

Algorithm

Step 1 : Select the *K* initial mode value

Step 2 : For dividing the clusters the following objective function is used

$$d_c(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j)$$

and

$$\delta(x_i, y_j) = \begin{cases} 0, x_j = y_j \\ 1, x_j \neq y_j \end{cases}$$

Step 3 : Choose an object to the correct cluster using the mode value

Step 4 : Analysis the dissimilarity of the object against current mode.

Step 5 : Repeat the steps 2,3 and 4 till the last tuple in the dataset.

B. Squeezer Algorithm

In this algorithm the data are clustered using the attribute value and their corresponding support value. The algorithm produces high quality of clusters with deserves good scalability [18].

Algorithm

Step 1 : Select the first tuple in the given dataset

Step 2 : Generate the cluster structure

Step 3 : Compute the similarity using the following support measure

$$Sim(C, tid) = \sum_{i=1}^{m} \frac{\sup(a_i)}{\sum_j \sup(a_i)}$$

Step 4 : In this step the data are clustered using the threshold value. If the threshold value in high the given data is assigned to the new cluster or else it assigns the data to the existing cluster

Step 5 : Repeat the steps for 2 to 4 until the end of the tuple in the given dataset.

C. ROCK Algorithm

It is an agglomerative hierarchy clustering algorithm. It measure the link similarities among the data points. This algorithm is divided into two stages. In the first stage, each tuples are assigned to each cluster. In the second stage, clusters are merged based on the closeness between the clusters [19].

Algorithm

Step 1 : Select a random samples from the given dataset

Step 2 : The samples are measured using the link similarity

Step 3 : Based on the link similarities the data are clustered using the following criterion

$$E_l = \sum_{i=1}^{m} n_i * \sum_{p_q, pr \in c_i} \frac{link(p_q p_r)}{n_i^{1+2f(\emptyset)}}$$

Step 4 : Finally the clusters are labeled

D. K-HISTOGRAM

This algorithm is the extended version of the K-means algorithm. In this algorithm the mean value are replaced by the histograms during the data clustering process [20].

Algorithm

Step 1 : Initialize the *K* value

Step 2 : Use the cost function given in the following equation

$$P(W,X) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_i l^d (X_i H_l)$$

Step 3 : The data objects are clustered based on the nearest histogram value

Step 4 : The histogram values are updated after the each assignment.

Step 5 : Repeat the steps till all the data are assigned to the particular cluster

E. DBCSAN

The DBSCAN algorithm [21] clusters the data using the nearest density based data points. This algorithm requires two important parameters. The first parameter is maximum radius neighbor data points (Eps). The second parameter is (MinPts) minimum number of points in the E_{PS} neighbor data points

Algorithm

Step 1 : Select a data point *p*

Step 2 : Select all data point from p satisfying the Eps

and MinPts Measure

Step 3 : If p is an relevant data point the selected data points is assigned to the cluster or else in assign to the existing cluster

Step 4 : The above steps are repeated till all the tuples are clustered in the give dataset

F. CURE Algorithm

The Cure algorithm [22] cluster the data with the fixed number of data points. Which are represented in the scatter matrix. The scatter matrix are shrinked to find the nearest data point to group the data into a cluster. In thi algorithm the outliers are efficiently detected.

Algorithm

Step 1 : Select the random sample

Step 2 : Based on the samples the data points are portioned.

Step 3 : In this step, the data points are clustered partially.

Step 4 : The outliers are detected in the given dataset and are eliminated.

Step 5 : The above steps repeated till the end of the data points.

V. CONCLUSION

In this paper, a survey is carried out to know how the data are clustered using the feature selection techniques. Many feature subset selection proposed by many researchers were. In section 3 various statistical measures were discussed and this measure is used to cluster the data in an efficient manner. The main aim of the existing algorithms is to find the optimal feature subset. But, finding the optimal feature subset from the unstructured data still remains difficult. The future work is to propose a new algorithm to find the optimal feature set to improve the clustering accuracy with less computational cost.

REFERENCES

[1] M.Vijayalakshmi, M.Renuka Devi, "A Survey of Different Issue of Different clustering Algorithms Used in Large Data sets" , International Journal of Advanced Research in Computer Scienceand Software Engineering, pp.305-307, 2012.

[2] Huan Liu,Hiroshi Motoda, Rudy Setiono and Zheng Zhao, "Feature Selection: An Ever Evolving Frontier in Data Mining",Journal of Machine Learning Research, volume 10, 2013, Hyderabad, pp. 4- 13, ISSN: 1938-7228

[3] Seoung Bum Kim, Panaya Rattakorn, "Unsupervised feature selection using weighted principal components", International journal of Expert Systems with Applications, Volume 38 Issue 5, May, 2011, pp 5704- 5710.

[4] M. Ramaswami and R. Bhaskaran, "A Study on Feature SelectionTechniques in Educational Data Mining", Journal of Computing Volume 1, Issue 1, December 2009, pp.7-11,ISSN: 2151-9617.

[5] Khedkar S.A., Bainwad A. M., Chitnis P. O, A Survey on Clustered Feature Selection Algorithms for High Dimensional Data, International Journal of Computer Science and Information Technologies, Volume 5, Issue 3, 2014,pp. 3274-3280

[6] M. Dash, K. Choi, P. Scheuermann, and H. Liu, Feature Selection for Clustering – A Filter Solution, Proceedings of the Second International Conference on Data Mining, 2012, pp. 115–122.

[7] Duy-Dinh Le, and Shin'ichi Satoh, An Efficient Feature Selection Method for Object Detection, Springer LNCS, Volume 3686, 2012, pp. 461-468

[8] Fengxi Song, Zhongwei Guo, and Dayong Mei, Feature Selection Using Hybrid Feature Selection Techniques, Proceedings of the IEEE International Conference on System Science, Engineering Design and Manufacturing Informatization, Volume 1, 2010, pp. 27-30.

[9] Y. Kim, W. Street, and F. Menczer, Feature Selection for Unsupervised Learning via Evolutionary Search, Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, pp. 365–369.

[10] Alan Agresti, "*An Introduction to categorical data analysis*", Wiley Series in Probability and Statistics, Second Edition, Wiley-Interscience, 2010.

[11] A. Rajaraman and J. D. Ullman, "Mining of Massive Datasets", Cambridge University Press, 2011.

[12] A. Desai, H. Singh and V. Pudi, "DISC: Data Intensive Similarity Measure for Categorical Data", Proceedings of Advances in Knowledge Discovery and Data Mining – 15th Pacific Asia Conference, Vol. 6635, 2013, pp. 469 – 481,

[13] D. Ienco, R. G. Pensa and R. Meo, "From Context to Distance: Learning Dissimilarity for Categorical Data Clustering", *ACM Transactions on nowledge Discovery from Data*, Vol. 6, No. 1, 2012.

[14] Xiaofei He, Ming Ji, Chiyuan Zhang, Hujun Bao, A Variance Minimization Criterion to Feature Selection Using Laplacian Regularization, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 33, Number 10, October 2011, pp. 2013-2025.

[15] Feiping Nie, Shiming Xiang,Yangqing Jia, Changshui Zhang, Shuicheng Yan, Trace ratio criterion for feature selection, 23rd National Conference on Artificial Intelligence, Volume 2, 2008, pp. 671-676, ISBN: 978-1-57735-368-3

[16] A. Gretton, O. Bousquet, A. Smola, B. Schoelkopf, Measuring Statistical Dependence with Hilbert-Schmidt Norms, Proceding of the 16th International Conference Algorithmic Learning Theory, 2010, pp. 63-78. DOI:270036

[17] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical

Values", Data Mining and Knowledge Discovery, Vol. 2, No. 3, 2009, pp. 283 – 304.

[18] R. Ranjani, S. A. Elavarasi and J. Akilandeswari, "Categorical Data Clustering using Cosine based similarity for Enhancing the Accuracy of Squeezer Algorithm", International Journal of Computer Applications, Vol. 45, No. 20, pp. 41-45, 2012.

[19] S. Guha, R. Rastogi and K. Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes",Information Systems, Vol. 25, No. 5, 2010, pp. 345 – 366.

[20] Z. He, X. Xu, S. Deng and B. Dong, "K-Histograms: An Efficient Clustering Algorithm", International Journal of Data mining, Volume 20,2009,pp 25-30

[21] M. Ester, H. P. Kriegel, J. Sander and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases", *International Conference on Knowledge Discovery and Data Mining*,2009, pp. 222 – 231.

[22] S. Guha, R. Rastogi and K. Shim, "CURE: An Efficient Clustering Algorithm for Large Databases", Proceedings of the ACM SIGMOD International Conference on Management of Data, 2012, pp. 73 – 84.