IJCTA, 10(9), 2017, pp. 575-582 ISSN: 0974-5572

CBICA: Correlation Based Incremental Clustering Algorithm

Kaustubh Shinde* and Dr. Preeti Mulay**

ABSTRACT

With progress in the area of computer science, it is achievable to read, process, store and generate information out of the available data. Humongous amount of data is generated, which is of mixed type, including time-series, Boolean, spatial-temporal and alpha-numeric data. This data is generated at a giant speed and volume, which makes difficult for the traditional clustering algorithms to create and maintain the desired clusters. Thus, the proposed system encourages incremental clustering using a non-probability based similarity measure. The experimental results, of Correlation Based Incremental Clustering Algorithm (CBICA), which are obtained using the Pearson's coefficient of correlation, are compared with the experimental results of the Closeness-Factor Based Algorithm (CFBA), which uses the probability based similarity measures. The threshold computation is done to decide the cluster members in the post clustering phase, to adapt influx of new data. Wherein the new data is accommodated in the available clusters or new clusters are formed, depending upon the threshold values.

Index Terms: Closeness-Factor, Cluster, Correlation, Incremental Clustering.

1. INTRODUCTION

In the last few decades, digitization is taking a boom. Various new sensors are used for the digitization process. These sensors generate a massive amount of data, which is of mixed type, including time-series, Boolean, spatial-temporal and alpha-numeric data. This data is generated at a giant speed and volume, which makes tough for the conventional clustering algorithms to create and maintain the desired clusters. The available incremental clustering algorithms are intended to cluster the influx of new data, but cannot handle the order insensitive data. The proposed system is designed in such a way which uses the non-probability similarity measures, which is Pearson's coefficient of correlation.

Various types of datasets which are handled by the proposed system are:

- Numeric datasets (Time-series and Boolean).
- Mix-type datasets (Spatial-temporal and Alpha-numeric).

These datasets, used as raw data, are input to the system CBICA. Cleansing of raw data is done by some of the data pre-processing techniques.

Few of the data pre-processing techniques are:

- Weight assignment.
- Removing zeros.
- Removing duplicates.
- Feature extraction.
- Principle Component Analysis (PCA).

* Symbiosis Institute of Technology, *Email: kaustubh.shinde@sitpune.edu.in*

** Symbiosis Institute of Technology, *Email: preeti.mulay@sitpune.edu.in*

Once the data pre-processing techniques are used on the raw data, the incremental clustering algorithm CBICA is used on the raw data, to generate the basic clusters, depending upon the closeness value. This closeness value is calculated using a set of formulas, which includes the computation of correlation, weight and error. The basic clusters which are generated using this closeness value are stored in the cluster database.

The principle motivation for CBICA system was to obtain a *pure incremental clustering*, *order insensitive* and *scalable* algorithm. Incremental clustering algorithms are the intelligent algorithms which incrementally clusters the influx of new data. After arrival of the new data to the system, again the closeness is calculated and depending upon its values, either the existing clusters are updated or new clusters are generated. These updated or generated clusters are stored in the cluster database. The patterns of stored clusters are compared with the patterns of clusters obtained from the reordered i.e. order insensitive input data. If the patterns match, then we can conclude that CBICA is an *order insensitive* algorithm.

In the post-clustering phase, some purity-based validation techniques, such as Precision, Recall, F-measure, are used to validate the clusters. Comparison between the outputs of probability based and non-probability based similarity measures is done. The clusters with PCA and without PCA are compared as well. The consistency of an incremental clustering algorithm is proved only when it combines two data series into the same cluster which generates the same distribution [1].

2. LITERATURE REVIEW

Researchers have proposed various incremental clustering algorithms over the period of time, but these algorithms have some limitations. To overcome these, the proposed system clusters the various data sets available.

2.1. Multi-Assignment Clustering for Boolean Data

This research paper achieves a novel approach, called Multi-Assignment Clustering (MAC), for clustering Boolean data that can concurrently be part of several clusters [2]. An expectation-maximization (EM) algorithm is presented where the source framework of the clusters and the cluster memberships of each data item are concurrently estimated. Average Hamming Distance is the post validation technique used, which suggest how precisely the centroids of the underlying clusters are evaluated. The proposed system does not support the concept of data to be simultaneously belonging to multiple clusters, hopping of data is not allowed, which is not effective for the efficiency of the incremental clustering algorithm.

2.2. Evolve systems using incremental clustering approach

This research paper presents a new paradigm for machine learning and recommends a new structure for incremental clustering. CFBA attained all the attributes necessary for a ‘pure incremental clustering’ and almost ‘parameter-free’ algorithm [3]. Scaling of dynamically available data is comfortably achieved using CFBA. CFBA converges comfortably being ‘cluster-first’ approach and not ‘centre-first’. This CFBA is used to enhance the proposed system, by comparing and proving accuracy of the results.

2.3. Knowledge augmentation via incremental clustering: new technology for effective knowledge management

This research paper attains a new incremental clustering algorithm rooted on closeness, a systematic and scalable approach which updates clusters and learns new information effectively [4]. The alpha-numeric data, which was used, was first transformed into numeric data, using the weight-assignment algorithm. This transformed numeric data is given as input to the Closeness Factor-Based Algorithm (CFBA). Confusion matrix is used to validate the results, but its drawback is that, CFBA works only for the numeric data and thus this can be thought of as a beneficial line of future research. Focusing on this point, the proposed system is made efficient to work for mix-type of data as well, of any domain.

2.4. Effective Clustering of Time-Series Data Using FCM

This research paper employs a two-level fuzzy clustering approach to obtain the dimensionality reduction [5]. Upon dimensionality reduction by a symbolic presentation, time-series data are clustered in a high level phase by making use of the longest common subsequence as similarity measurement. Then, by making use of an effective method, framework is made based on generated clusters and progressed to the next level to be reused as initial centroids. Afterwards, a fuzzy clustering approach is used to explain the clusters precisely. This concept is used in the proposed system, when the incremental clustering is to be done. Here the prototypes are referred as the basic clusters which are generated initially, and then further used to update or append the clusters after influx of new data.

2.5. Threshold Computation to Discover Cluster Structure, a New Approach

This research paper achieves a comparative study about shaping cluster structure and is based on selecting an optimal threshold value to form a cluster [6]. Incremental clustering approach “Incremental clustering using Naïve Bays and Closeness-Factor” (ICNBCF). Using Naïve Bays, the threshold computation is done. The concept of threshold computation is used to form the clusters depending upon the closeness values calculated. The threshold value varies from domain to domain.

Considering the work done related to the incremental clustering algorithms, the possible enhancement is to replace the probability based similarity measure with the non-probability based similarity measures. The most appropriate non-probability based similarity measure is the Pearson’s coefficient of correlation.

Table 1
Comparison between probability and correlation.

Sr. No.	*Probability*	*Correlation*
1	Probability here is the ration between each series(sample) and the grand total(population).	Correlation will give us the relation between two series.
2	It is an unbiased approach.	It is an biased approach.
3	The range is 0 to +1.	The range is -1 to +1.
4	Only positive relation can be obtained.	Both positive as well as negative relation is obtained.

Table [I] gives us the idea, why probability should be replaced by correlation. The range of correlation gives the positive as well as the negative relation between the data series, which is an added advantage over the probability which only gives us the positive relation between the data series. The correlation has a biased approach, which gives equal importance to all the data series in the input data.

3. PROPOSED MODELS OF CBICA

3.1. Mind Map of CBICA

The Mind Map consists of the following sub-sections:

1. Input data of various types.
2. Pre-clustering step.
3. Clustering step.
4. Incremental clustering step.
5. Post-clustering step.

In fig. [1], input to the system is the raw data sets of various types. The non-numeric data is converted to numeric data using weight assignment algorithm. The cleansing of data by performing either of the following; Remove noise, PCA, Remove zeros, Remove duplicates.

In the initial clustering step, a batch of data is processed by computing sum of attributes, coefficient of correlation, error, closeness and threshold. Using these computed values, the basic clusters are formed, which are updated in the cluster database. The series_processed_flag are set to true if the series belongs to a cluster.

Incremental clustering is performed using the influx of new data or reordered data. The computation of different values is done to update the existing clusters or to update the existing clusters or the new clusters. The clusters obtained from new data or reordered data are compared to check whether the same pattern is followed or not.

In the post-clustering step, some of the techniques used to validate the clusters are Precision, Recall and F-measure. Comparison between the probability based and non-probability based similarity measures is

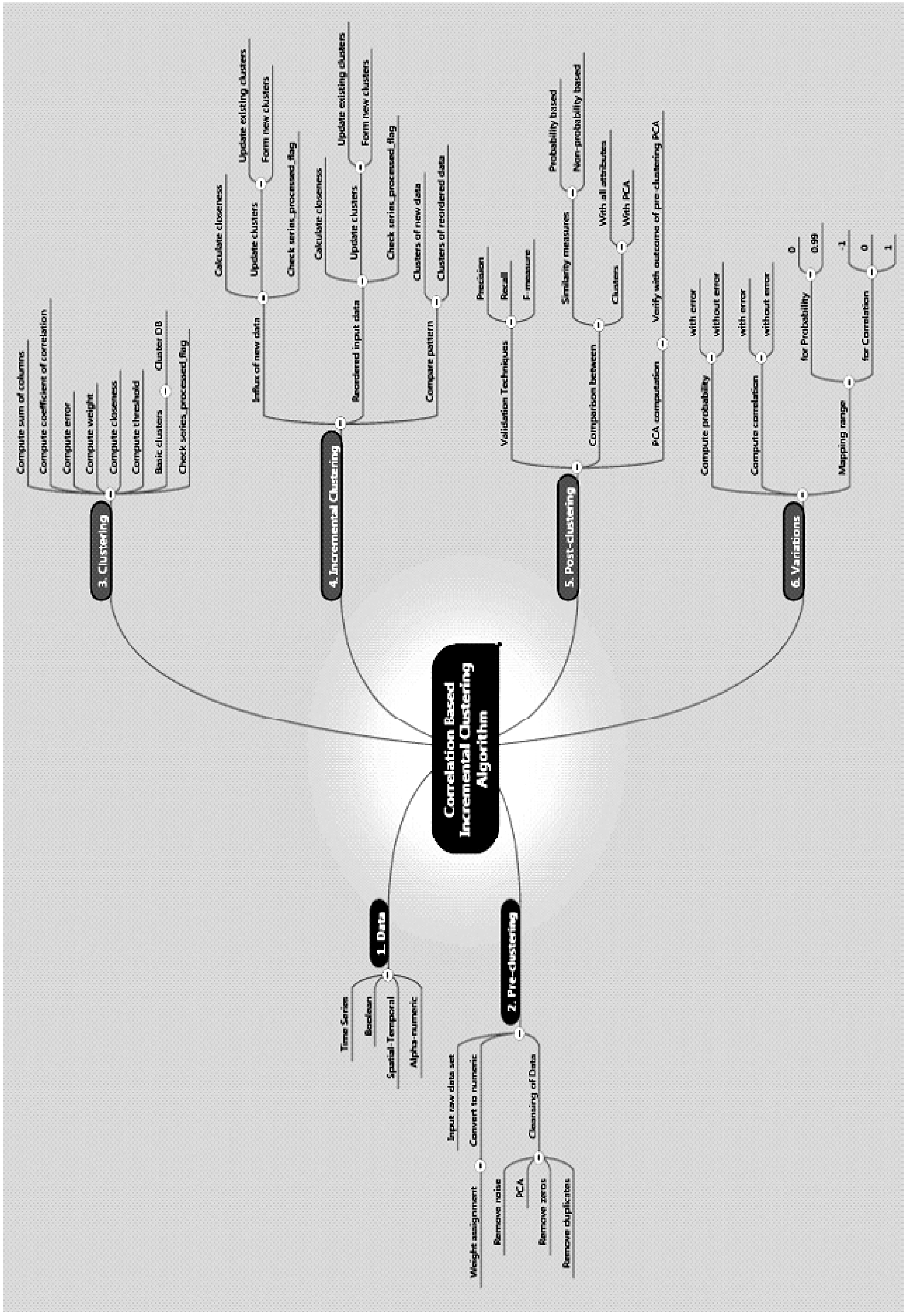

Figure 1: Mind map of CBICA.

carried out along with the comparison of clusters which are derived using all attributes and which are derived using only the PCA. From these final clusters, the PCA are obtained and verified along with the PCA obtained from the pre-clustering.

The variations are computing probability and correlation values with and without considering error.

3.2. Architecture of CBICA

The CBICA is designed in such a way that, it can be used to incrementally cluster various datasets available.

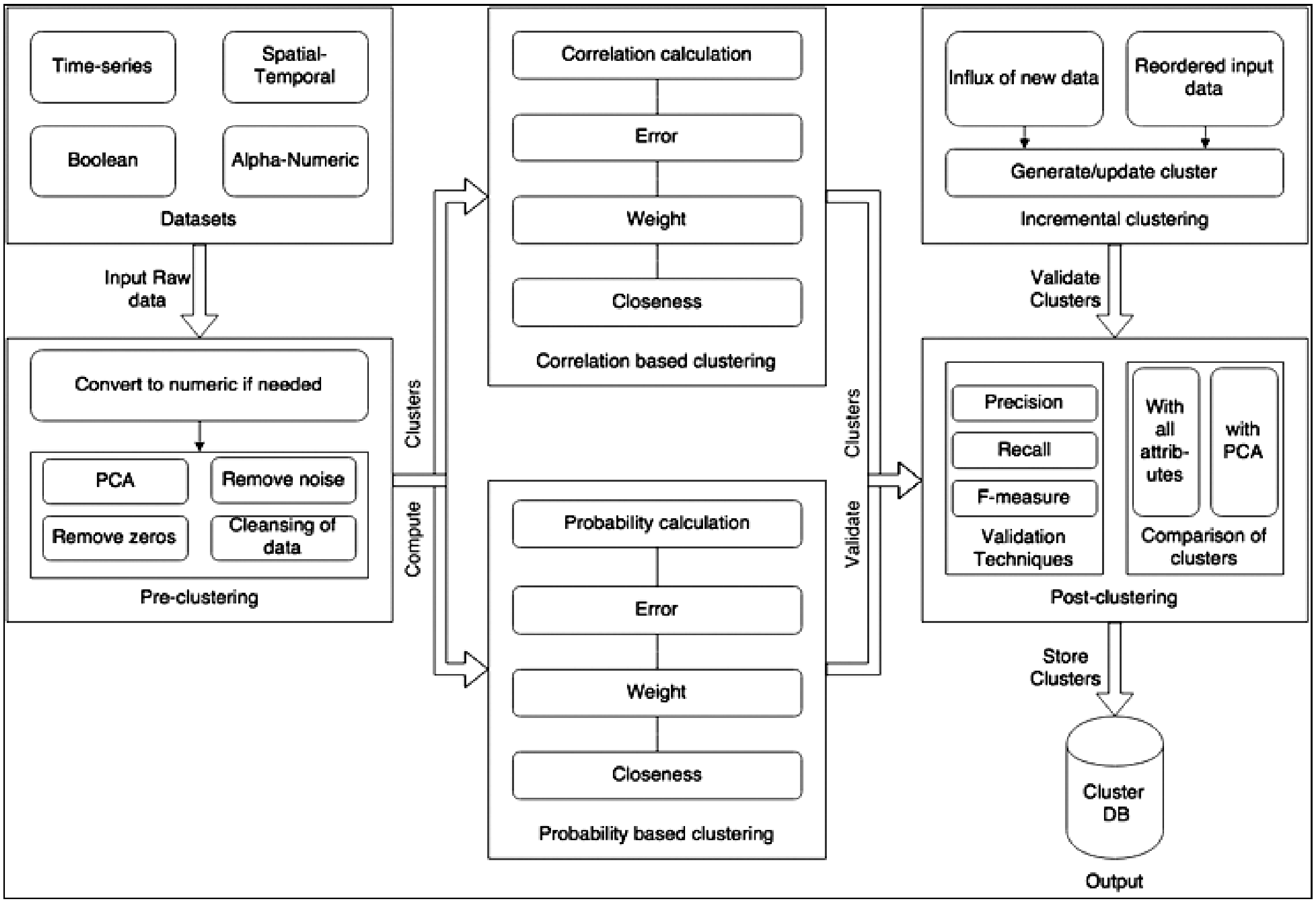

Figure 2: Architecture of CBICA.

In fig. [2], initially a dataset, amongst the various available datasets, is taken as a raw data, as an input to the system. This input raw data is converted into the numeric data, using the weight-assignment algorithm. Further the pre-processing of data is done, by any of the pre-processing techniques available. In the next step, CBICA is applied to the pre-processed data. As this algorithm uses a “cluster-first” approach, user is not responsible for defining the number of clusters. The basic clusters are generated initially. These clusters are saved into the cluster database. Due to the “cluster-first” approach, hoping of cluster members is not possible. If there is an influx of new data, the incremental approach can be used. The incremental approach uses all the above steps to update the existing clusters or generates new clusters if needed. These updated or newly generated clusters are also saved into the cluster database.

The next important step is to validate the clusters, in the post-clustering phase. Statistical validation techniques are used to validate the numeric data; whereas conceptual validation techniques are used to validate the categorical data. Various validation techniques are used to validate the clusters such as variance, Dunn index, F-measure and Rand index.

The variance and the Dunn index are the internal validation techniques, whereas the F-measure and Rand index are the external validation techniques.

4. IMPLEMENTATION OF CBICA

The initial implementation part was done using the Wine dataset from the UCI Repository [7]. The dataset characteristics are multivariate. It consists of 178 instances and 13 attributes. The initial implementation consists of the computation of the series total, probability, error, weight, closeness value and the how-close value, which is used to obtain the patterns of each cluster. The probability is the ratio between the series total and the grand total. Following are some of the results obtained from the how-close values of the generated clusters.

Fig. [3] and fig. [4] are the patterns of the basic clusters C1 and C2, generated from the *how-close* values of the first batch of input data. The *how-close* value must be under 1.0, or else the series will be considered as an outlier. The cluster C1, in fig. [3], consists of 18 cluster members, whose closeness values range between 0-0.99.

The cluster C2, in fig. [4], consists of 12 cluster members, whose closeness values range between 1-1.99.

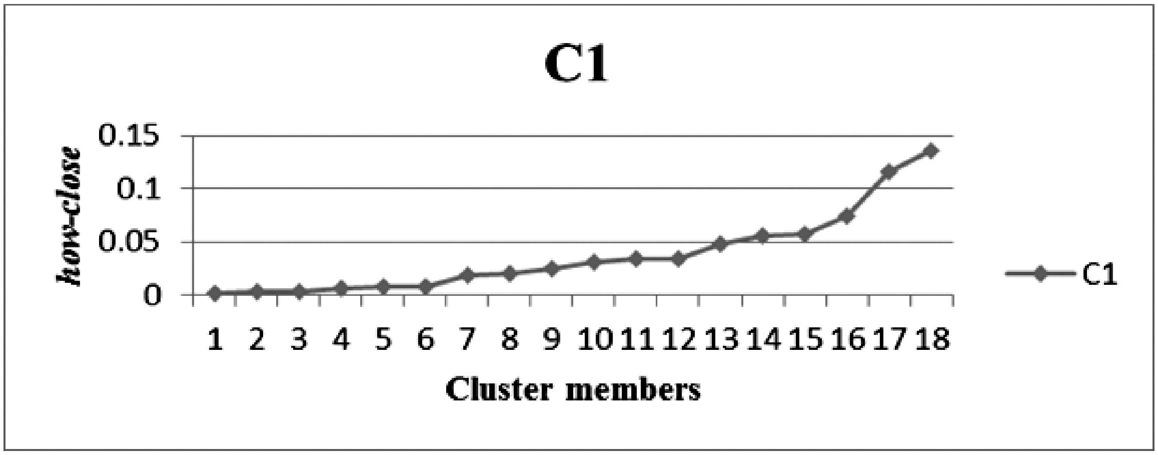

Figure 3: Pattern of cluster C1.

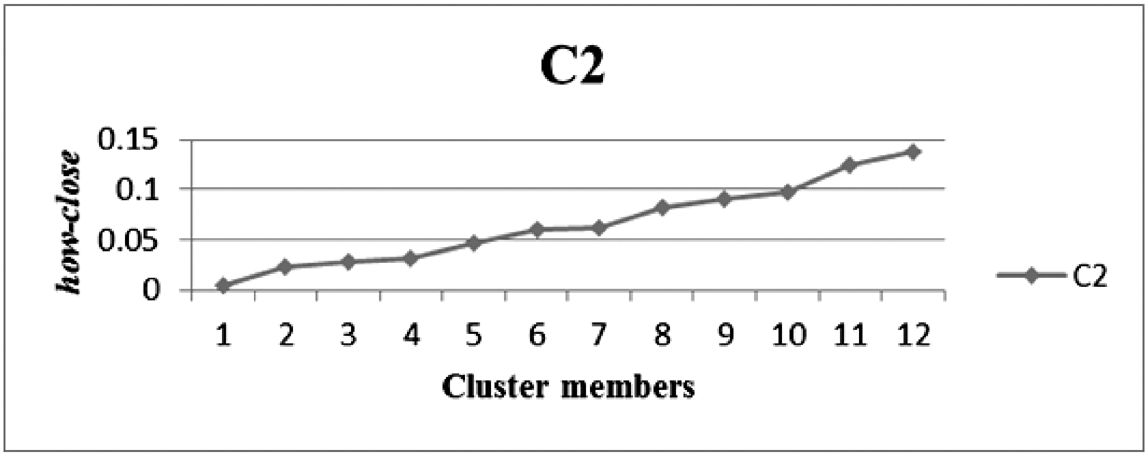

Figure 4: Pattern of cluster C2.

Fig. [5] and fig. [6] are the patterns of the clusters NC1 and NC2, generated from the *how-close* values of the new batch of input data. The cluster NC1, in fig. [5], consists of 10 cluster members, whose closeness values range between 0-0.99. The cluster NC2, in fig. [6], consists of 5 cluster members, whose closeness values range between 1-1.99. The clusters NC1 and NC2, are the new clusters obtained from the influx of new data. Once the patterns are obtained of the initial and new clusters, the incremental part to merge the

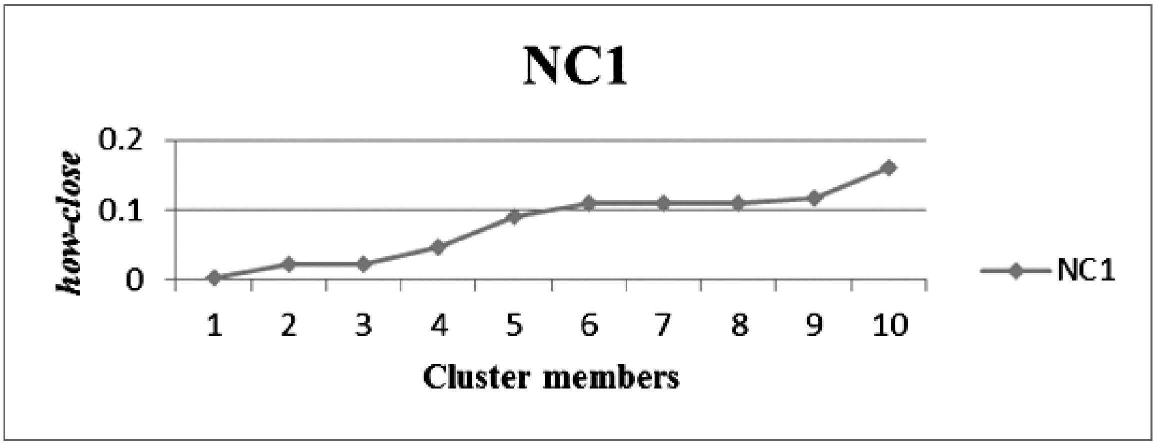

Figure 5: Pattern of cluster NC1.

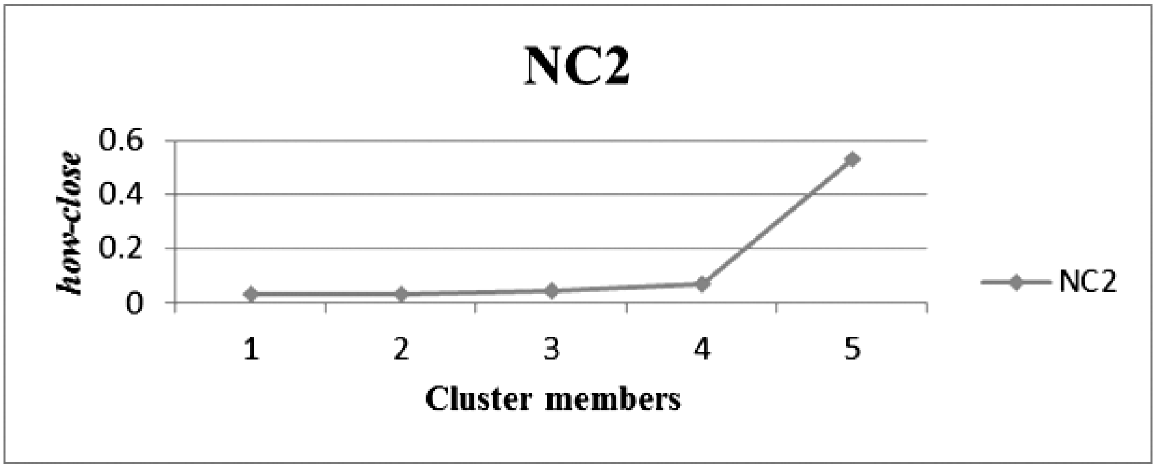

Figure 6: Pattern of cluster NC2.

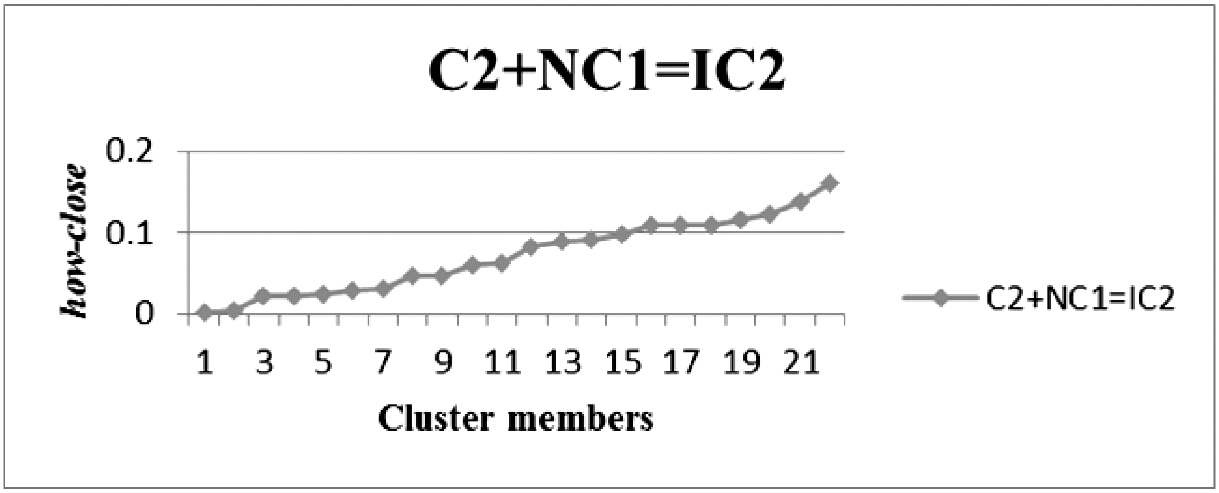

Figure 7: Pattern of cluster IC2.

clusters by mapping the patterns, takes place. Fig. [7] is the representation of the incremental cluster IC1 after the clusters C2 and NC1 are merged.

5. SUMMARY

The purpose of this paper is to review the trends and existing work related to incremental clustering algorithms. This review highlights the various incremental clustering algorithms and the datasets upon which they are used. This review helps to propose a system which overcomes the drawbacks of the existing incremental clustering algorithms, which uses the probability based similarity measures. The proposed system uses non-probability based similarity measure, i.e. Pearson’s coefficient of correlation, and shows promising results over the probability based incremental clustering algorithm. As the range of correlation is between -1 to +1, the outcomes clearly suggests whether the two data series are negatively correlated or positively correlated. This is an added advantage over the probability based incremental clustering algorithm. The proposed system further needs to calculate the threshold values, which are used to assign the input data to the desired clusters, is part of future work.

REFERENCES

[1] Khaleghi, A., Ryabko, D., Mary, J. and Preux, P., “Consistent algorithms for clustering time series,” Journal of Machine Learning Research, Vol. 17, No. 3, pp.1-32, 2016.

[2] Frank, M., Streich, A.P., Basin, D., and Buhmann, J.M., “Multi-Assignment Clustering for Boolean Data,” Journal of Machine Learning Research, Vol. 13, pp. 459-489, 2012.

[3] Kulkarni, P.A., and Mulay, P., “Evolving Systems using incremental clustering approach,” Evolving Systems, Vol. 4, pp. 7-85, DOI:10.1007/s12530-012-9068-z, 2013.

[4] Mulay, P., and Kulkarni, P.A., “Knowledge augmentation via incremental clustering: new technology for effective knowledge management,” Int. J. Business Information Systems, Vol. 12, No. 1, pp.68–87, 2013.

[5] Aghabozorgi,S., and Wah,T.Y., “Effective Clustering of Time-Series Data Using FCM,” International Journal of Machine Learning and Computing, Vol. 4, No. 2, 2014.

[6] Mulay, P., “Threshold Computation to Discover Cluster Structure, a New Approach,” International Journal of Electrical and Computer Engineering, Vol. 6, No. 1, pp. 275-282, 2016.

[7] https://archive.ics.uci.edu/ml/datasets/Wine