A Novel Approach for Clustering based on Pattern Analysis

Prachi M. Joshi
Research Scholar
Department of Computer Engineering and Information Technology, College of Engineering, Pune, India

Dr. Parag A. Kulkarni
Adjunct Professor
Department of Computer Engineering and Information Technology, College of Engineering, Pune, India

ABSTRACT
Clustering aims at grouping of data into clusters based on the similarity between them. It is the pattern of the data that governs grouping. In this paper, we propose method for clustering that is based on finding closeness between the data series. A novel method referred as Clustering with Closeness factor (CCF) is proposed that works in two phases and is not pre-bound with clusters numbers. The method identifies the pattern of data and performs clustering. With proper selection of threshold value, the approach can prove to be a big step for decision making.

General Terms
Machine intelligence, machine learning and pattern analysis

Keywords
Clustering, closeness, threshold, k-means

1. INTRODUCTION
The problem of classification is getting more and more critical with the availability of large amount of unlabelled data. A good clustering must be more compact in same group and more separate among different groups [1].

There are many approaches for clustering, but there is always a tradeoff between performance and accuracy. A problem that is commonly faced is deciding the optimal number of clusters for the data set. Understanding the pattern of data and then generating optimal clusters needs to be focused. Most of the methods for clustering require the number of clusters to be specified prior to clustering [2]. In practice, the other issue that needs attention is the number of scans the data has to go through at the same time generating robust clusters.

We present a novice clustering approach that reveals the patterns in the data and clusters them and is not pre-bound with the cluster number.

2. RELATED WORK
Clustering has taken place in wide range of applications starting from text and media categorization, intrusion detection to health care [3] [4] [5] [6]. In recent years, a number of clustering approaches have been proposed with varied application perspectives, each of them aiming to improve the overall clustering.

Among the clustering methods, the most common method used being the k-means [2]. K-means always needs the cluster number to be specified in advance and its performance is driven by the initial centroid. Methods with variants in k-means [7] [8] [9] [10] [11] show improvements in initial centroid calculation as well as consider the type of cluster formulation. The number of scans data undergoes before the clusters are stabilized is a factor that needs to be looked upon.

Hierarchical techniques tend to consider the data points that are local neighbors and the overall size of the clusters is unseen [12] [13]. It is necessary that the underlying pattern be identified and the clustering takes place in robust way.

Techniques generating cluster with the similarity measure are required to be robust [12] [14] in terms of the outliers, the volume and initialization. New techniques to build clusters also have evolved with methods of matrix factorization techniques to improve the clustering efficiency [15]. Besides these, methods are extended to have maximum margin separation between the formed clusters and probabilistic approaches towards the cluster formulation have engaged attention. [16] [17].To achieve better clustering performance; probabilistic methods with pattern analysis is what that interests researchers.

In this work, we consider the task of building a clustering approach to cluster unlabeled data set U_d. Information about the number of clusters formed is unknown. The algorithm on the basis on the data pattern analyzes and generates the clusters. Working in two phases, the approach takes into account the robust parameter of the cluster as well.

3. PROPOSED CLUSTERING FRAMEWORK
The proposed approach works in two phases. In the first phase, initial cluster set is built and then in the second phase, the clusters formed are restricted, to get robust cluster. The methodology does not use distance measure for determining the relative closeness between the series. It makes use of a novice approach - ‘closeness factor’. Baseline of the approach is finding the similar series to reveal the patterns and then generate the clusters. The method needs threshold value to be set which is used as limit value during the formulation of clusters. The clusters are generated dynamically and results show that with good threshold value the proposed algorithm has clustering properties that are worth to notice.

3.1 Closeness Value
The proposed closeness factor is discussed here. With this closeness, the decision about which series is close to other, when to generate the cluster, which series should be added to a cluster, all is dependent on this closeness value.

The closeness between the series is calculated with the probabilistic approach.

Consider two data series Sr_1 and Sr_2. $Sr_x(i)$ is point i in series x. Sum(i) is the total of the corresponding parameters of the series considered.

The probability of outcome Sr_1 is calculated as

$$P = \left[\sum_{i=1}^{n} Sr_1(i) \Big/ \sum_{i=1}^{n} Sum(i) \right]$$

The expected value of $Sr_x(i)$ is calculated

$$Sr_x(i) = P * Sum(i)$$

An error err(i) is defined as

$$err(i) = P * Sum(i) - Sr_x(i) \Big/ \sqrt{Sum(i) * P * (1-P)}$$

Finally, the closeness 'C' between the series is calculated as

$$C^2 = \sum_{j=1}^{n} err(i)^2 * wt(i) \Big/ \sum_{j=1}^{n} wt(i)$$

Where wt is the weight equals to $\sqrt{sum(i)}$

With the equation, the Closeness value is determined. The lower the value of closeness; the closer are the series.

3.2 Phase 1: Building initial cluster set

In the 1st phase of the approach, the series are compared with each other. The comparison is done with the closeness factor. The comparison is to understand the pattern of the data and generate the clusters. This needs threshold value for limiting and determining the cutoff for the clusters formed.

Given set of unlabeled data U_d and threshold ∂ value, the algorithm works as follows:

1. Read the data set.
2. Compute and compare closeness values.
3. With respect to the threshold ∂ and closeness, generate new cluster if not satisfying the criteria or append the data to existing one.

The decision criterion for building the clusters is detailed as:

Assume that P is a pivot series with which new series S is evaluated in terms of similarity. C is the closeness between P and S.

If($C<\partial$) and C<lowest-value; where lowest-value is earlier calculated closeness initialized to zero, then update the lowest closeness value at the same time-

Assign cluster to P and S based on following:

(i)If P already assigned, assign S to same cluster.

(ii)If S already assigned, assign P to same cluster.

(iii)If none are assigned already, generate new cluster and assign both to this cluster.

At the end of the phase one, labeled data set- LDS is generated. The labeled data generated in the phase one is further refined. This is done to avoid the misclassified data and get better accuracy. The clusters generated in the first phase are substantial in number. This labeled data goes as input to phase two; where robust clusters are generated.

3.3 Phase 2 of CCF

Re-clustering takes place in this phase. It is necessary that the clusters be apart and there is a substantial distance between them to take care of boundary conditions. For the generation of crisp sets, maintaining the accuracy and retaining the cluster feature, phase two is required. A new labeled data series LDS is generated that clusters the data with accuracy.

The labeled data set now is treated again to be unlabeled, to refine the clusters. The clusters formed at the end of phase one are grouped in sorted manner with respect to the labels generated. The decision criterion now is as follows:

Assume that S is already assigned to cluster x, now decision of pivot on closeness and ∂ is done as:

If $C<\partial$ and C<lowest-value earlier calculated, then update the lowest value and

(i) Assign pivot to cluster as that of S. Add pivot to LDS.

(ii) If the pivot does not satisfy the criteria, generate new cluster for it and add to LDS.

At the end of the second phase, we get the clusters generated that maintain the accuracy of the labeled data with formulation of crisp sets. No cluster number is specified in priori to the algorithm in any of the phases.

3.4 Threshold Value - ∂

Though the clusters generated are incremental in nature, the methodology needs proper selection of threshold value. The value is critical to consider the base level cut

With an accurate threshold value selection, accurate clusters are generated. It was found that average closeness value of the unlabeled data gave good accuracy for clustering and hence was selected to be the threshold criteria.

$$\partial = avg(closeness\_all)$$

4. EXPERIMENTS AND RESULTS

The proposed methodology was tested for wine, wine quality and the iris data sets from the UCI Machine Learning Repository [18].

Following graph shows the purity results for wine and iris data sets with CCF.

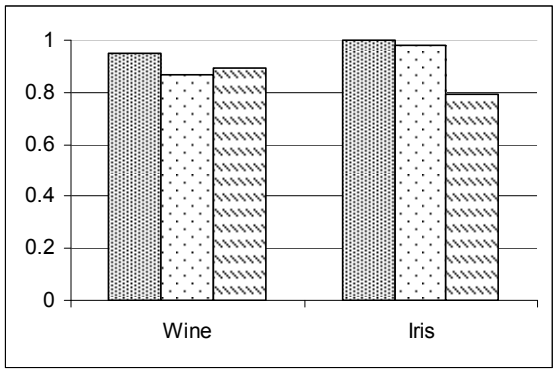

Fig. 1 Purity of formed clusters with CCF for wine and iris data

The proposed method is compared with k-means. The following table shows points of comparison between k-means and the proposed method.

Table 1: Comparison between k-means and CCF

Comparison Factor	K-means	Proposed Approach
Cluster Number	Required before clustering	Not required
Threshold	Not required	Required
Cluster Formulation dependency	Centroids selected	Closeness

From the above table, it is observed that CCF method overcomes the dependency on centroids selection and the cluster numbers specified in advance. CCF has tried to handle the data in a different approach based on pattern analysis to come up with a new clustering methodology.

Purity results for wine, wine quality and iris data set with k-means, hierarchical and CCF. The table below shows that CCF can prove to be better option in clustering.

Table 2: Purity results: Comparison with different methods

Approach	Wine	Wine quality	Iris
k-means	0.89	0.70	0.83
hierarchical	0.90	0.85	0.87
CCF	0.90	0.82	0.88

5. CONCLUSION AND FUTURE WORK

The results that we have here are with small cluster number and are just the start in order to address problem with larger magnitude. A novel approach for clustering with closeness is put forth. It is not just the threshold value but the dynamic change in closeness value that generated the clusters accurately. Extension of the work include investigating CCF approach for some more datasets as well as application of the approach as a baseline method for incremental update of clusters that can be applied in semi-supervised way.

6. REFERENCES

[1] Camastra, F. and Verri, A. 2002. A novel kernel method for clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, p 22-32.

[2] Kanung, T. Netanyahu, N. and Wu, A. 2002. An efficient k-means clustering algorithm, analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, p 881-892.

[3] Kulkarni, P., Kulkarni, M. 2002. Advance forecasting methods for resource management and medical decision-making. In Proceedings of National Conference on Logistics Management: Future Trends.

[4] Becker, H., Namman M., and Gravano, L. 2010. Learning similarity metrics for event identification in social media. In Proceedings of the Third ACM International Conference on Web Search and Data Mining, USA, p 291-300.

[5] Bharti, K., Jain, S., and Shukla, S. 2010. Fuzzy k-means clustering viaJ48 for intrusion detection system. International Journal of Computer Science and Information Technologies, Vol. 1, p 315-318.

[6] Lui, Y., Cai, J., Yin, J., and Fu, A. 2008. Clustering text data streams. Journal of Computer Science and Technology,(Jan 2008) , p 112-128.

[7] Deelers, S., Auwantanamongkol, S. 2007. Enhancing k-means algorithm with initial cluster centers derived from data partitioning along the data axis with highest variance. International Journal of Computer Science.

[8] Fahim, A., Saake, G., Salem, A., Torky, F., and Ramadam, M. 2008. K-means for spherical clusters with large variance in sizes. Journal of World Academy of Science, Engineering and Technology.

[9] Jain, A., Murthy, M., and Flynn, P. 1999. Data clustering: A review. ACM Computing Surveys, Vol. 31, No. 3.

[10] Lai, J., Huang, T., and Liaw, Y. 2009. A fast k-means clustering algorithm using cluster center displacement. Journal of Pattern Recognition, Vol. 41, No. 1, (Nov 2009), p 2551-2556.

[11] Chandra, E., Anuradha, V. 2011. A survey of clustering algorithms for data in spatial database management systems. International Journal of Computer Applications, Vol. 24, No. 9.

[12] Dave, R., and Krishnapuram, R. 1997. Robust clustering methods: A unified view. IEEE transactions on fuzzy systems, Vol. 5, No. 2.

[13] Geva, A. 1999. Hierarchical unsupervised fuzzy clustering. IEEE Transactions on Fuzzy Systems, Vol. 7, No. 6.

[14] Yank, M., and Wu, K. 2004. A similarity based robust clustering method. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 26, No. 4.

[15] Cai, D., He, X., and Han, J. 2011. Locally consistent concept factorization for document clustering. IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 6, 902 – 913.

[16] He, Q., Chang, K., Lim E., and Banerjee, A. 2009. Keep it simple with time: A Re-examination of probabilistic topic detection models. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 10, No. 10.

[17] Zhang, K., Tsang, I., and Kwok, J. 2009. Maximum margin clustering made practical. IEEE Transactions on Neural Networks, Vol. 20, No. 4, (April 2009).

[18] Frank, A., Asuncion, A. 2010. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.