Essence-Based Clustering: A Multi-Strategic and Highly-Customizable Clustering Approach

Kevin Christian Rodríguez-Siu, Dennis Barrios-Aranibar, Raquel E. Patiño-Escarcina
Grupo de Investigación en la Línea de Automatización Industrial, Robótica y Visión Computacional - LARVIC
Centro de Investigación e Innovación en Ciencia de la Computación - CICC
Universidad Católica San Pablo - Arequipa, Perú
Email: {kcrodriguez,dbarrios,rpatino}@ucsp.edu.pe

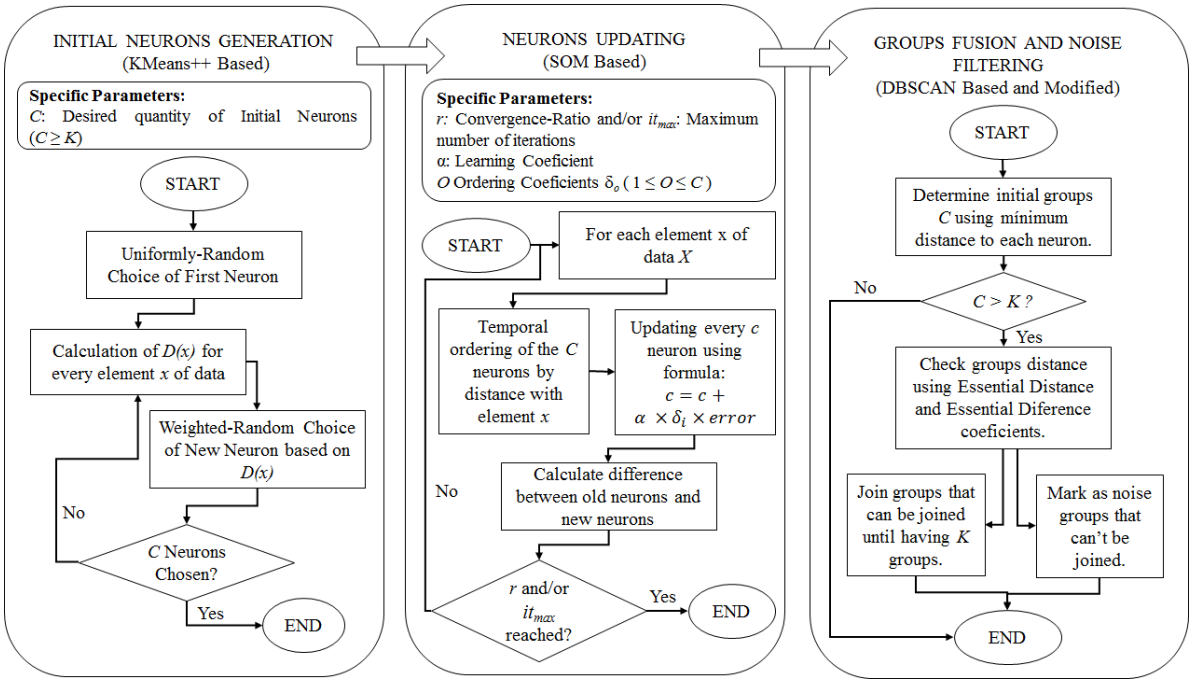

Fig. 1: Blocks Diagram of the Algorithm

I. Multiple Strategy Clustering

The idea of mixing traditional clustering strategies to improve the final results is not something new. Many recent works try to combine different clustering approaches to get better results. To name some examples, we have works exposing a combination of two strategies: the K-Means++ method and the Self-Organizing Map (SOM) method in a new algorithm called SOM++, that greatly improves the results each method can obtain on its own. Older approaches tried to combine also the KMeans algorithm with other methods, namely an Ant Colony Based SOM and an Adaptive Resonance Theory 2 (ART2) Neural Network.

More recent approaches also try to combine other AI techniques with already known solutions for clustering, like combining an Artificial Immune System algorithm and K-means, or methods based on K-modes with fuzzy genetic algorithms. The main issue with all of these approaches is that they propose a full combination of the algorithms, making them more complex to implement and understand. The merge of different strategies then becomes another one-single strategy that can't be broken down in steps easily to optimize each process individually, when due to their nature they should be broken down in more customizable steps.

TABLE I: Iris and Soybean Data Set: Evaluation Results

Iris	EM	Spectral C.	EBC-25	EBC-50	EBC-100
Silhouette Coef.	**0.6857**	0.5442	0.4895	0.4986	0.5605
Ad. Rand Score.	0.5584	0.7592	0.5822	0.5634	**0.8032**
Ad. Mut. Info S.	0.551	0.7934	0.6034	0.5804	**0.8231**
Homogeneity S.	0.5537	0.796	0.6087	0.5860	**0.8253**
Completeness S.	0.9490	0.8156	0.9086	0.9311	**0.9515**
V. Measure S.	0.6994	0.8057	0.7252	0.7192	**0.8315**

Soybean	EM	Spectral C.	EBC-4	EBC-8	EBC-12
Silhouette Coef.	**0.4425**	0.3169	0.3141	0.3169	0.3041
Ad. Rand Score.	0.5949	0.5513	0.5512	**0.6888**	0.5513
Ad. Mut. Info S.	0.6662	0.6888	0.6888	**0.7055**	0.6888
Homogeneity S.	0.6842	0.7222	0.7206	**0.7296**	0.7222
Completeness S.	**0.9306**	0.7138	0.7138	0.7295	0.7138
V. Measure S.	0.7886	0.718	0.718	**0.7296**	0.718

TABLE II: Wine and Breast Cancer (WDBC) Data Set: Evaluation Results

Wine	EM	Spectral C.	EBC-3	EBC-6	EBC-12
Silhouette Coef.	0.4357	0.5587	0.4969	**0.5612**	0.1249
Ad. Rand Score.	0.0289	0.3542	**0.4186**	0.3636	0.0156
Ad. Mut. Info S.	0.0983	0.4041	0.4025	**0.409**	0.0999
Homogeneity S.	**0.7433**	0.4153	0.4117	0.4104	0.11
Completeness S.	0.2	0.4157	0.4088	**0.4722**	0.219
V. Measure S.	0.3152	0.4155	0.4103	**0.4391**	0.1464

WDBC	EM	Spectral C.	EBC-2	EBC-3	EBC-5
Silhouette Coef.	**0.7563**	0.4065	0.6643	0.698	0.6921
Ad. Rand Score.	0.0026	0.4195	**0.6062**	0.4555	0.5338
Ad. Mut. Info S.	0.028	0.4069	**0.5005**	0.3911	0.4579
Homogeneity S.	**1.0**	0.427	0.5012	0.3918	0.4587
Completeness S.	0.1111	0.4077	**0.5571**	0.4935	0.5448
V. Measure S.	0.1999	0.4171	**0.5277**	0.4368	0.498

II. Essence-Based Clustering (EBC)

The use of one strategy simplifies greatly the task of the performance of the algorithms. But it also limits the process as only one strategy usually can't correct wrong previous choices. This causes algorithms to fall into local optimizations, or to not consider alternative solutions. The idea of Essence-Based Clustering (or EBC for short) is to perform not one, but many strategies of clustering in the same algorithm with two main goals in each strategy used:

1) Optimize the grouping of the data formed in a previous process.
2) Be an alternative solution in case a previous or later process is discarded.

The inclusion of more strategies in the same algorithm changes the usual configurations of parameters, so while all of them can work together as a whole, each step of the algorithm has enough information to work on its own. Then, parameters of the algorithm are classified in two categories:

1) Generic Parameters, that affect all the algorithm processes.
2) Specific Parameters, that affect only one strategy or process of the algorithm.

This causes an algorithm that can be changed in two levels, as the processes can be affected globally (in case generic parameters are changed) or locally (in case specific parameters are changed). The algorithm that has been implemented to prove the EBC approach is a three-step process, where each step corresponds to one strategy used in traditional clustering. The whole process is schematized in Figure 1.

III. Testing

To evaluate the algorithm, specific metrics to evaluate clustering performance have been used. There are two different types of metrics: Internal Evaluation Measures, that are the metrics that evaluate the result of the clustering based on the data that was clustered itself; and External Evaluation Measures, where clustering results are evaluated based on known class labels and external benchmarks which consist of a set of pre-classified items, and these sets are often created by human experts. We used a mix of both to evaluate the EBC approach.

IV. Conclusions

This work has presented a new clustering algorithm which is based on a combination of three different existing strategies for tackling this problem. It has been found that, in general, the most traditional problems and spherical groups of data require a small number of initial neurons in our algorithm; but when the data set is larger or the groups are not spherical, a greater number of initial neurons is required to give a more accurate solution.