An Energy Exchanging Mechanism for Data Clustering

Roberto Alves Gueleri and Liang Zhao

Institute of Mathematics and Computer Science
University of São Paulo
São Carlos – São Paulo – Brazil
gueleri@icmc.usp.br; zhao@icmc.usp.br

Abstract—In this paper, a dynamic process for data clustering is presented. It is based on the collective behavior among the objects of the input dataset. Each object is assigned an energy state, so they interact with each other by exchanging their energy, causing similar objects to take similar states. Finally, a classical algorithm such as k-means is applied on the energy vectors to actually cluster the data. Experiments show that the energy exchanging process is able to transform complex arrangements of objects into arrangements much easier to cluster. Moreover, the energy exchanging process is resilient to the mixture of clusters to some extent.

Keywords—clustering; collective behavior; emergence; self-organization; swarm intelligence.

I. INTRODUCTION

Data clustering (or just clustering) plays a key role in the field of Machine Learning. Informally, the clustering problem can be stated as: given a set of unclassified objects, the goal is to find the best partition of this set, such that objects belonging to the same cluster are more similar than those belonging to different clusters [1]–[4]. Despite this simple (yet informal) definition, clustering is a complex, challenging task [1]–[4], so there are still many efforts on devising more powerful clustering algorithms.

Collective behavior is a class of phenomena in which the actions of an individual is strongly influenced by the actions of each other [5]–[8]. A very interesting feature of these phenomena is that, even in systems consisting of simple interactions among simple individuals, these interactions may spontaneously lead to the emergence of a global pattern [5]–[8].

Some approaches based on collective behavior have been proposed for the clustering task. Blatt et al. [9] proposed a clustering algorithm based on the physical properties of an inhomogeneous ferromagnet. Each object is assigned a spin. Then each object interacts with its neighbors and the clusters appear as objects whose spins become aligned. Zhao et al. [10] proposed a clustering algorithm based on attraction of similar objects. Similarity is relative to the distance between them. So this attraction drives the objects within a certain region to converge to their corresponding cluster center. Abraham et al. [11] and Oliveira et al. [12] show the application of ant colony and particle swarm optimization for data clustering. In some of these approaches, each individual represents a possible solution and the interactions among them lead to a near-optimal solution.

In this paper we present an energy exchanging mechanism that is able to drive the clustering process. It is a dynamic system based on the collective behavior among its individuals. Each object from the dataset constitutes an individual of the system and is assigned an energy state. So they interact with each other by exchanging their energy, and a sort of self-organization emerges from this system, causing similar objects to take similar energy states. Then we apply k-means algorithm [3], [4], [13], [14] on these energy states in order to actually group the objects.

The rest of this paper is organized as follows. A formal description of the energy exchanging mechanism is given in Sec. II. Experiments over artificial datasets are given in Sec. III. Finally, Sec. IV concludes the paper.

II. PROCESS DESCRIPTION

Here we formally describe the energy exchanging mechanism. Initially, each object is assigned an energy state. As the process advances, these objects progressively exchange their energy. The distance between each pair of objects defines their degree of interaction. This interaction decreases as the distance becomes longer, being positive for close objects and negative for distant ones. Positive interaction causes the energy states to get closer, while negative interaction, in turn, moves the energy states away from each other. It is expected that, for a proper choice of parameters, the system reaches a stationary state at which the objects from a same class maintain their energy states closer than the objects from different classes.

Strictly speaking, this process does not realize the complete data clustering task. Instead, it maps the objects from the original space into a new one, called energy space. Then the objects in this new energy space are supposed easier to be correctly grouped.

Let $n$ be the number of attributes of each data item. So each object $o_i$ is given by the vector

$$o_i = (o_{i,1}, o_{i,2}, \ldots, o_{i,n})$$  \hspace{1cm} (1)

where each component $o_{i,k}$ is a real-valued attribute.

Let $n$ be the number of objects, and $t$ the iteration index of the dynamical system. Thus, each object $o_i$ is linked to
an energy state vector \( e_i \),

\[
e_i(t) = (e_{i,1}(t), e_{i,2}(t), \ldots, e_{i,n}(t))
\]

(2)

where each component \( e_{i,k}(t) \) represents the real-valued energy level of the object \( i \) at the iteration \( t \). Notice that, unlike the objects in the original space, their energy states are time-dependent, so they move through the energy space.

An interaction degree \( \text{inter}(o_i, o_j) \) is defined for each pair of objects,

\[
\text{inter}(o_i, o_j) = \text{pos}(\|o_i - o_j\|) - \text{neg}(\|o_i - o_j\|)
\]

(3)

where \( \|o_i - o_j\| \) is the Euclidean distance, and the portions \( \text{pos}(\|o_i - o_j\|) \) and \( \text{neg}(\|o_i - o_j\|) \) are functions such that \( \text{pos}(\|o_i - o_j\|) \) is strictly decreasing, \( \text{neg}(\|o_i - o_j\|) \) is strictly increasing, and both are positive.

Thus, the resultant interaction function is strictly decreasing and results zero for some positive distance. This is in accordance to the idea that the farther the objects are, the lower is their interaction. Also, this function is positive for small distances and negative for greater ones. If the distance between some pair of objects results in a zero-interaction, then there won’t be any energy exchange directly between them.

The system evolution is then governed by

\[
e_{i,k}(0) = \begin{cases} 0 & \text{for } i \neq k \\ 1 & \text{for } i = k \end{cases}
\]

(4)

\[
e_{i,k}(t + 1) = e_{i,k}(t) + \alpha \sum_{o_j} (e_{j,k}(t) - e_{i,k}(t)) \text{inter}(o_i, o_j)
\]

(5)

where \( o_j \) stands for all the objects (not including \( o_i \)) such that

\[
\text{inter}(o_i, o_j) < 0 \implies (e_{i,k}(t) > 0 \land e_{j,k}(t) > 0)
\]

(6)

that is, if \( \text{inter}(o_i, o_j) < 0 \), then neither \( e_{i,k}(t) \leq 0 \) nor \( e_{j,k}(t) \leq 0 \); otherwise there will be no exchange between \( o_i \) and \( o_j \) for the energy component \( k \). This constraint puts a lower bound of zero for any component of energy. Actually, due to the discrete nature of this process, these components may take values just below zero.

The constant \( \alpha \) determines the precision of this process. Smaller values lead to greater precision, but make the convergence slower. Greater values, in turn, can even make the entire process unstable. A good choice for \( \alpha \) should balance precision and convergence time. This choice is affected by the number and arrangement of objects in the dataset, and by the interaction function as well. Therefore, it is not a trivial task to choose a proper value for \( \alpha \), and this may even be an empirical choice.

From (5), one can see that for each gain there is a loss of energy. So the sum of all energy remains constant throughout the process. Precisely,

\[
\sum_i e_{i,k}(t) = 1 \quad \forall k, t
\]

(7)

that is, for any iteration \( t \), the sum of any component \( k \) over all the objects \( i \) equals 1.

For positive interaction, this process looks somewhat like a physical heat conduction. Energy is transferred from high-energized objects to lower ones. Also, the transfer rate is proportional to the difference between their energy, as can be seen from (5). As this difference decreases, the exchange becomes slower. If there was only positive interaction among the objects in the dataset, then the whole set would be led to a single energy state. For negative interaction, however, the system acts the opposite way, forcing the objects to take energy states more and more different from each other, recalling that the energy components can only take values between 0 and 1. An illustration of how these two different interactions (positive and negative) act upon the energy states is given in Fig. 1. It is important to remark that, despite this analogy to physical systems, we are not intending to bring real physical concepts to our energy exchanging model.

As an ending note, given a dataset and a set of parameters, this process is entirely deterministic. Each energy state vector demands as many components as the number of objects in the dataset. As stated in (4), this high dimensionality allows for a total independence among these energy states at the beginning. As a consequence, since this process does not rely on any random choice for the initial state, no object can eventually take advantage (or disadvantage) of these arbitrary choices. Thus, the cluster formation process is completely determined by the dataset itself, together a set of parameters.

### III. Experiments

Here we present two sets of experiments performed over two different artificial datasets. The general idea is, firstly, to
run the dynamic process described in the previous section. Then \textit{k-means} algorithm, one of the most popular clustering algorithms [3], [4], [13], [14], is applied on the resultant set of energy states. Finally, the quality of the obtained partitions is measured. For this purpose, the \textit{Rand index} is used [3], [4]. It is an external cluster validation measure and, as such, compares the outcome of an algorithm to a given reference partition [3], [4].

All experiments employed the following class of interaction function:

$$\text{pos}(d) = pd^{-\beta}$$

(8)

$$\text{neg}(d) = nd^{1/4}$$

(9)

yielding

$$\text{inter}(o_i, o_j) = \text{pos}(d) - \text{neg}(d) = pd^{-\beta} - nd^{1/4}$$

(10)

where \(d\) denotes the distance between the objects \(o_i\) and \(o_j\), that is, \(d = ||o_i - o_j||\), and \(p\), \(n\), \(\beta\) and \(\gamma\) are positive constants taken as parameters. It is desirable that we can control the distance at which the interaction results zero. Let’s call this distance just \textit{zero at} \(\alpha\), and we can set it by properly adjusting the parameters \(p\) and \(n\). An example of how this class of interaction function behaves is shown in Fig. 2. It has \(\beta = 2\), \(\gamma = 4\) and value zero for \(d = 5\).

All experiments were carried out through the following steps:

1) Choosing values for \(\alpha\), \(\beta\) and \(\gamma\), as well as the desired distance for which the interaction results zero. Actually, proper values for \(\alpha\) were obtained empirically;
2) Starting the process and letting it run up to a desired number of iterations. We chose 3000 iterations;
3) Running \textit{k-means} algorithm over the resultant set of energy vectors. Then getting the Rand index over the resultant partition. We employed a version of \textit{k-means} that sets initial centroids at different randomly chosen objects [3], [4], [13], [14]. Since \textit{k-means} is sensitive to its random initialization and is prone to get trapped in very poor local minima [3], [4], [13], [14], we actually ran \textit{k-means} 50 times, then we averaged the 10 best results.

A. \textit{Experiment 1 – A Ring Around}

The results of Experiment 1 are presented in Fig. 3. It was performed on the dataset shown in Fig. 3(a), which is composed of two classes: one whose objects are scattered just around the center (Class 1), and the other whose objects are scattered over a circumference (Class 2). Class 1 contains 100 objects and Class 2 contains 300 objects.

Results of different combinations of \(\beta\) and \(\gamma\) are shown in Figs. 3(c), (d) and (e). The Rand index from a mere random assignment of those objects over two clusters and by assigning all the objects to only one cluster are 0.5 and 0.624, respectively. We bring the latter because the dataset is unbalanced. Notice that the maximum value of Rand index is 1, meaning the resultant partition exactly matches the reference partition [3], [4]. On the other hand, applying \textit{k-means} on the original dataset is meaningless, since \textit{k-means} cannot cope with this sort of arrangement, so it would end up with very poor results [3], [4], [13], [14].

Values for \(\alpha\) were chosen as shown in Table I. For instance, for \(\beta = 1\), it ranges from 0.00025 (zero at 1) to 0.00005 (zero at 4). Intermediate values are a linear interpolation. For practical reasons, however, at the very beginning of the process, \(\alpha\) was set to a very small value and progressively increased up to its right value. This is because the process appears to be too violent at the beginning, and a smaller value for \(\alpha\) softens it.

Consider one of the best results, namely, one of that for \(\beta = 3\), \(\gamma = 2\) and zero at 2.2, which results in Rand index 0.966. The evolution of this run is shown in Fig. 3(f). There are six measurements taken throughout the process: quartile 1 (Q1), median and quartile 3 (Q3) for the distances between all the pairs of objects (in terms of their energy states), divided into pairs of the same class (intra-class) and different classes (inter-class). One can see that the process appears to be reasonably stable after some time and that the classes appear to get somewhat separated. Finally, in Fig. 3(b) it is shown the resultant clusters.

From this experiment, we see that there is some range of parameters, specially for \(\beta = 3\), leading to very good results. If we look at the original dataset, we notice there are a few objects of Class 1 entirely mixed with Class 2.
Figure 3. The results of Experiment 1. (a): Original dataset. (b): One of the best results, namely, one of that for $\beta = 3$, $\gamma = 2$ and zero at 2.2. (c), (d) and (e): Results of different combinations of $\beta$ and $\gamma$. (f): Evolution of the process for $\beta = 3$, $\gamma = 2$ and zero at 2.2. There are the quartile 1 (Q1), median and quartile 3 (Q3) of the distances between the energy states of all pairs of objects, divided into pairs of the same class (intra-class) and different classes (inter-class).
So it becomes very hard to correctly group these objects, what decreases the index value. From Fig. 3, we see that k-means acts quite well on the energy vectors transformed by the proposed mechanism, in contrast to the original dataset. These results suggest that the energy exchanging mechanism is suitable for datasets whose classes draw a dense region of objects and are apart from each other by sparser regions.

B. Experiment 2 – Different Sizes

The results of Experiment 2 are presented in Fig. 4. It was performed on the dataset shown in Fig. 4(a). This dataset is composed of two classes scattered around two points. The first class (Class 1) has 80 objects. The other (Class 2) has 320 objects and covers a wider area. Although each of these classes (alone) have a shape suitable for k-means (objects scattered around a point) [3], [4], [13], [14], their different sizes (in terms of area) make it a difficult task.

As in the previous experiment, the results of different combinations of $\beta$ and $\gamma$ are shown in Figs. 4(c), (d) and (e). There is a comparison with the Rand index on a random guess, and on a partition whose objects all belong to a single cluster, this latter resulting 0.679. There is also the Rand index by applying k-means on the original dataset, which results about 0.874. Like the processed dataset, the results from original one are the average of the best 10 of 50 runs.

Parameter $\alpha$ took exactly the same values as in Experiment 1. The evolution of one of the best results and its resultant clusters are shown in Figs. 4(f) and (b), respectively. This result is one of that for $\beta = 3$, $\gamma = 2$ and zero at 2.8, which results in Rand index 0.980.

Like the previous experiment, we see very good results, but now for a broader range of parameters. Again, k-means acts better on the new energy space than on the original dataset. This dataset imposes a further mixture at the boundary between the classes. Still, the energy exchanging mechanism seems to be resilient to this higher mixture.

IV. CONCLUDING REMARKS

As we see throughout the experiments, the energy exchanging mechanism is suitable for some datasets whose classes are apart from each other by low-density regions. Also, this process is resilient to the mixture of classes to some extent. It drives the objects so that a somewhat complex arrangement becomes easier to a classical clustering algorithm such as k-means.

In the process presented in this paper, each energy state vector demands as many components as the number of objects in the dataset. As previously mentioned, whereas this high dimensionality requires more computational resources, this leads to a cluster formation process that is completely determined by the dataset itself. So the results can be analyzed without the influence of some random choices. Currently, the authors are engaged in a study that aims to measure how much reducing that dimensionality impacts on the quality of the resultant clusters. Also, they are studying the impact of an interaction that occurs not between all the pairs of objects, but instead between a (possibly random) subset of that pairs.

ACKNOWLEDGMENT

This work is supported by the São Paulo Research Foundation (FAPESP), by the Coordination for the Improvement of Higher Education Personnel (CAPES), and by the Brazilian National Research Council (CNPq).

REFERENCES

Figure 4. The results of Experiment 2. (a): Original dataset. (b): One of the best results, namely, one of that for $\beta = 3$, $\gamma = 2$ and zero at 2.8. (c), (d) and (e): Results of different combinations of $\beta$ and $\gamma$. (f): Evolution of the process for $\beta = 3$, $\gamma = 2$ and zero at 2.8. There are the quartile 1 (Q1), median and quartile 3 (Q3) of the distances between the energy states of all pairs of objects, divided into pairs of the same class (intra-class) and different classes (inter-class).