

FREQUENTLY ASKED QUESTIONS ABOUT HYPERBALL ALGORITHMS

Q1. Should HBL algorithm be classified as a clustering algorithm?

Directly quoting from [1]: "**Clustering** is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure. Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. "

"Besides the term *data clustering* (or just *clustering*), there are a number of terms with similar meanings, including *cluster analysis*, *automatic classification*, *numerical taxonomy*, *botryology* and *typological analysis*."

In the papers [2], [3] the word "category" is used instead of "cluster" but in essence the Hyperball Algorithms (HAs) perform the tasks of data / pattern clustering, recognition and classification. A "category" is a cluster of patterns exhibiting similar traits, according to a particular biologically inspired metric. The use of the word "category" rather than "cluster" is intended to emphasize the fact that this is a family of algorithms useful in performing three different (but related) computational tasks: clustering, recognition and classification. More reasons for introducing this distinction are given below. Traditional clustering algorithms were created to perform one task only: that of clustering.

Q2. If so, is HBL a new algorithm considering the previous clustering algorithms?

Yes. The Hyperball Algorithms emerged to address the inadequacy of previous clustering algorithms in mimicking the animal skills of pattern clustering, recognition and classification. Simple animals (insects, snails, etc.) do not have sufficient computing power to execute sophisticated algorithms proposed by humans, but still are capable of pattern clustering, recognition and classification. Frequently these animals have vision systems superior to humans (both in resolution and colour sensitivity) and execute their algorithms in real time to ensure their survival.

In particular, below are some deficiencies of previous clustering algorithms:

1. Some form of Minkowski metric is used, to measure the distance of points in sample space. This metric is denoted by the formula

$$d_k(p, q) = [|x_p - x_q|^k + |y_p - y_q|^k]^{1/k}$$

where p and q are points in sample space, x_i and y_i are their coordinates and $k \geq 1$ is a parameter, becomes ever more computationally expensive as the value of k increases. Observe that for $k=1$ this metric reduces to Manhattan metric, and for $k=2$ it becomes Euclidean metric. The HAs use the least expensive Manhattan metric, although are not limited to that metric.

2. In pattern clustering, recognition and classification we are clustering patterns rather than individual points in sample space. The word "cluster" applies to a set of data points being similar in some way (i.e. exhibiting a particular trait). The word "category" used by me applies to a set of patterns being similar in some way, where each pattern consists of a number of data points. No two patterns compared need to contain the same number of data points, hence the reason for using the term "category".

Traditional clustering algorithms organize sample data points into clusters. HAs organize patterns into categories. HAs therefore use the novel concept of distance D between patterns P and Q , denoted $D(P, Q)$, which can be derived from an arbitrary Minkowski metric, indeed, from any arbitrary metric between points p and q .

3. Traditional clustering methods leave the choice of metric to the user. In HAs the choice of preferred metric (minimizing computational cost) emerges from a biologically inspired reasoning, although arbitrary metric could be used, while the capability of clustering patterns rather than

single data points would remain.

4. Traditional clustering methods used in machine learning do not lend themselves easily to continuous learning, while for HAs continuous learning is a preferred methodology, although HAs could be used in traditional machine learning + testing regime.
5. Traditional clustering methods used in machine learning can themselves be clustered into two groups: those of supervised learning and autonomous learning. In the supervised learning methods the existence of an infallible expert teacher (i.e. some kind of "god") is presumed. HAs allow learning from a fallible expert (typically "parent"), who can commit mistakes, albeit sufficiently rarely. In this scenario the expertise of the learner can exceed that of a teacher. Additionally, novel teacher-learner collaborations can exist, improving expertise of both. Similar collaborations between multiple learners can also exist.
6. In traditional clustering methods the discernibility information (i.e. the information of maximum distance between two points to be included in the same cluster) is provided by the user, and then the system is "let loose" to cluster the data provided. Trouble with this approach is that the learning process is highly sensitive to that information, and wrong information leads to wrong learning. There is no general rule on what kind of discernibility information is suitable to any clustering job at hand.
7. Other traditional methods (k-means and its derivatives, like fuzzy c-means, etc.) are highly sensitive to the value of k to be provided at the beginning of learning. Trouble is, there are no guidelines as to proper choice of k, and wrong choice leads to wrong learning results. In HAs the discernibility information emerges through biologically inspired learning from errors.
8. QT clustering algorithm (QT = quality threshold) can figure the proper value k on its own, but is more computationally expensive than k-means.
9. Clustering methods relying on locality-sensitive hashing use Jaccard distance, which is highly expensive computationally, in a way inaccessible to simple animals.
10. Traditional clustering methods focus on the correctness of their main task of clustering first. The issue of performance takes a back stage. Hyperball algorithms, as all algorithms designed to perform in real time, have been designed with performance in mind, and are massively parallel to ensure real time response.

References:

- [1] http://en.wikipedia.org/wiki/Data_clustering
- [2] <http://www.cosc.brocku.ca/files/downloads/research/cs0807.pdf>
- [3] <http://www.cosc.brocku.ca/files/downloads/research/cs0808.pdf>