ForestFire: Label Propagation Clustering Inspired by Fire Dynamics

Jeremy Goldwasser

Advised by Prof. Mark Gerstein, Spring 2021

Introduction

Clustering is an unsupervised learning technique that partitions a dataset into groups of similar points. Most existing algorithms are limited by the assumptions they make about the data - for example, the number of clusters, their shape, and the amount of space between them. Our method, in contrast, requires minimal prior assumptions, using only one effective parameter. Per its name, ForestFire "burns" each cluster by selecting a centroid and iteratively propagating its label outwards. This framework is capable of learning a manifold of any shape, including on graphical data. Moreover, its outputs can be internally validated with a Monte Carlo-like simulation that evaluates the stability of each point's predicted label. We demonstrate that it achieves comparable or improved accuracy and runtime to 11 algorithms on 6 benchmarks, including Louvain, K-Means, and DBSCAN. We also show its utility on single-cell datasets, emphasizing its ability to identify rare cell types and doublet artifacts.

Methods

Our method formulates clustering as a forest fire, where points represent trees, clusters represent groves, and labels represent fires. A fire starts when a single tree, the flint, catches fire. The fire propagates to its neighbors, then to the neighbors of those points, and so on. When no new points are hot enough to ignite, we reiterate with a new flint. This process continues until all of the points have been assigned to a cluster. Lastly, we reassign all the points in miniscule clusters to larger ones.

To start a fire at the center of a cluster, ForestFire selects flints that are distanced from other clusters and in a dense area of the dataset. To do this, an affinity matrix is necessary. We calculate this matrix using a Gaussian kernel whose bandwidth parameter is roughly the distance between nearby points in the same cluster. Once a fire is lit, we iteratively calculate the heat at each point, also using the affinity matrix. These points ignite if their affinity to the fire is high enough, i.e. they are close enough to it.

Datasets

We benchmarked ForestFire on 6 toy datasets from Scikit-learn. We also applied it to 3 single-cell sequencing datasets, which contain genomic measurements of mouse brains. We chose single-cell data because clustering only works with algorithms that do not input the true number of clusters. This is due to measurements of unexpected classes like rare cell types and artifactual doublets.

Results

Figure 1 demonstrates that ForestFire achieved state-of-the-art performance on the clustering benchmarks, even when K is unknown. It achieved ≥98% accuracy on all but one toy dataset, something that only Louvain and Spectral Clustering achieved. Louvain, however, is incredibly slow, and Spectral Clustering requires K to be known. Moreover, neither is online or can provide a pointwise stability metric. Table 1 shows a longer list of advantages of ForestFire over other methods.

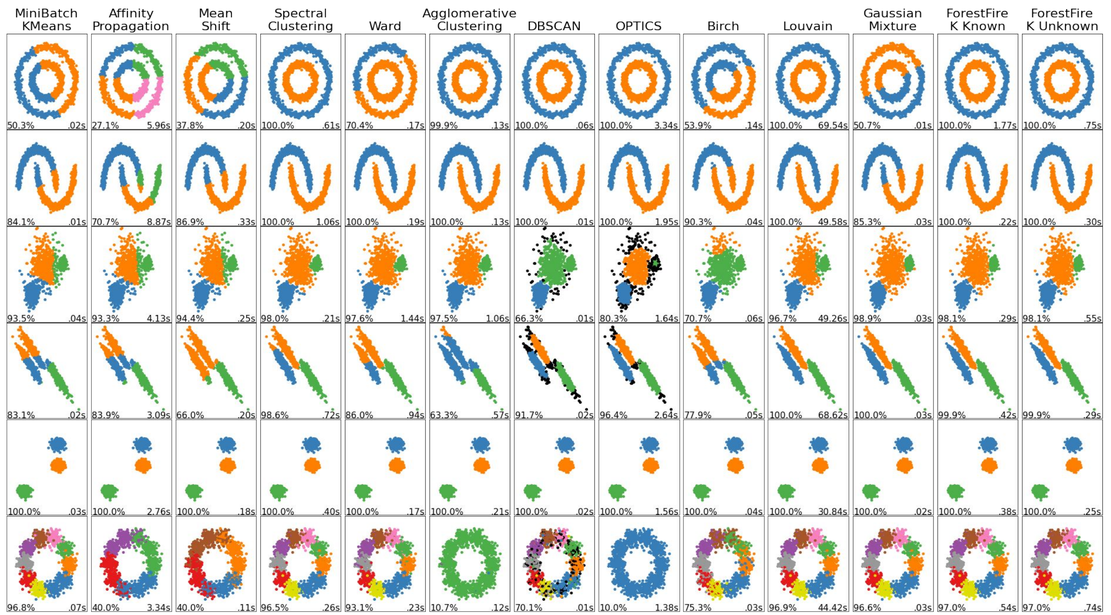

Figure 1: Benchmarking ForestFire against Existing Methods

On the dataset of 10 blobs, we ran 100 Monte Carlo simulations and calculated the entropy of their labels (Fig. 2). The results show, as one would expect, that points get progressively unstable the closer they get to a cluster boundary.

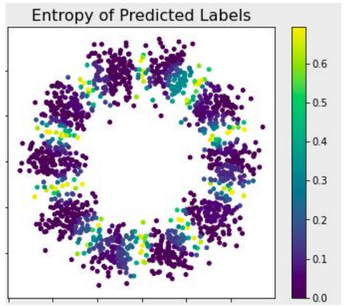

Figure 2: Stability Values

ForestFire also performs well on all three single-cell datasets (Fig. 3). These datasets had posed a particular challenge, as each one contained dozens of clusters.

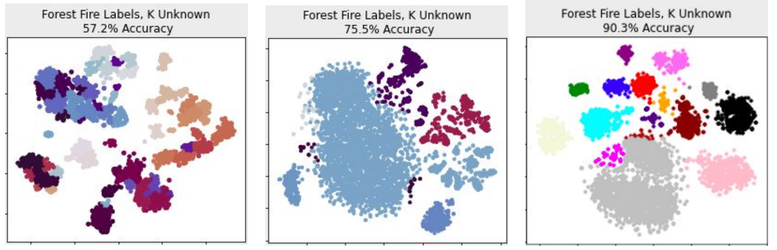

Figure 3: ForestFire applied to single-cell data

	ForestFire	Spectral Clustering	Louvain	K-Means	Gaussian Mixture	DBSCAN
Infers # of clusters	✅	❌	✅	❌	❌	✅
Learns manifold	✅	✅	✅	❌	❌	✅
Can add new data	✅	❌	❌	✅	✅	✅
Pointwise stability measure	✅	❌	❌	✅	✅	❌
Fast runtime	✅	✅	❌	✅	✅	✅
Handles adjacent clusters	✅	✅	✅	✅	✅	❌

Table 1: Feature Comparison

Conclusion

My thesis introduces ForestFire, a novel clustering algorithm inspired by forest fire dynamics. Unlike other algorithms, ForestFire requires minimal prior assumptions on the shape of the data to perform clustering. It can also use Monte Carlo simulations to produce pointwise confidence scores for its clustering results. We show that it performs as well or better than 11 leading algorithms on 6 benchmarks, exhibiting no significant drawbacks. Its flexibility and interpretability are useful in domains like single-cell analysis, where datasets may have more than the expected number of clusters.