Go Up

<table>
<thead>
<tr>
<th>Name</th>
<th>DBSCAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version</td>
<td>1.0.0</td>
</tr>
<tr>
<td>Description</td>
<td>DBSCAN Clustering Method</td>
</tr>
<tr>
<td>License</td>
<td><a href="http://www.apache.org/licenses/LICENSE-2.0">http://www.apache.org/licenses/LICENSE-2.0</a></td>
</tr>
<tr>
<td>Copyright</td>
<td>Copyright (C) 2019 HPCC Systems</td>
</tr>
<tr>
<td>Authors</td>
<td>HPCCSystems</td>
</tr>
<tr>
<td>DependsOn</td>
<td>ML_Core 3.2.2</td>
</tr>
<tr>
<td>Platform</td>
<td>7.4.0</td>
</tr>
</tbody>
</table>

Table of Contents

<table>
<thead>
<tr>
<th>DBSCAN.ecl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scalable Parallel DBSCAN Clustering Algorithm Implementation based on [1]</td>
</tr>
</tbody>
</table>

| DBSCAN_Types.ecl |

PARAMETER eps  || REAL8 — the maximum distance threshold to be considered as a neighbor of the other. Default value is 0.0.

PARAMETER minPts  || UNSIGNED4 — the minimum number of points required for a point to become a core point. Default value is 2.
PARAMETER dist ||| STRING — a string describing the distance metrics used to calculate the distance between a pair of points. Default value is 'euclidean'. Other supported distance metrics include 'cosine', 'haversine', 'chebyshev', 'manhattan', 'minkowski'.

PARAMETER dist_params ||| SET (REAL8) — a set of parameters for distance metrics that need extra setup. Default value is [] which should fit for most cases.

Children

1. fit : Fit function performs DBSCAN clustering on a dataset (ds) to find clusters and the cluster index (Label) of each sample in the dataset

2. Num_Clusters : Num_Clusters Provides the number of clusters that the given dataset will be divided into when clustered by the DBSCAN algorithm

3. Num_Outliers : Num_Outliers Provides the number of outliers that the given dataset will have when clustered by the DBSCAN algorithm

FIT fit

DBSCAN \n
| DATASET(ML_Core.Types.ClusterLabels) | fit |
| (DATASET(Types.NumericField) ds) |

Fit function performs DBSCAN clustering on a dataset (ds) to find clusters and the cluster index (Label) of each sample in the dataset.

PARAMETER ds ||| TABLE ( NumericField ) — The dataset in NumericField format to be clustered.

RETURN TABLE ( { UNSIGNED2 wi , UNSIGNED8 id , UNSIGNED8 label } ) — result in ML_Core.Types.ClusterLabels format describing the cluster index of each sample.

SEE ML_Core.Types.NumericField, ML_Core.Types.ClusterLabels
**NUM_CLUSTERS** Num_Clusters

DBSCAN

<table>
<thead>
<tr>
<th>DATASET(Files.l_num_clusters)</th>
<th>Num_Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>(DATASET(ML_Core.Types.ClusterLabels) ds)</td>
<td></td>
</tr>
</tbody>
</table>

Num_Clusters Provides the number of clusters that the given dataset will be divided into when clustered by the DBSCAN algorithm.

**PARAMETER** ds ||| TABLE ( ClusterLabels ) — A dataset with cluster index information. Usually it’s the result of Fit function.

**RETURN** TABLE ( { UNSIGNED4 wi , UNSIGNED4 num } ) — DATASET(l_num_clusters)
The number of clusters, per work item.

**SEE** DBSCAN_Types.l_num_clusters

---

**NUM_OUTLIERS** Num_Outliers

DBSCAN

<table>
<thead>
<tr>
<th>DATASET(Files.l_num_clusters)</th>
<th>Num_Outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>(DATASET(ML_Core.Types.ClusterLabels) ds)</td>
<td></td>
</tr>
</tbody>
</table>

Num_Outliers Provides the number of outliers that the given dataset will have when clustered by the DBSCAN algorithm.

**PARAMETER** ds ||| TABLE ( ClusterLabels ) — A dataset with cluster index information. Usually it’s the result of Fit function.

**RETURN** TABLE ( { UNSIGNED4 wi , UNSIGNED4 num } ) — DATASET(l_num_clusters)
The number of outliers, per work item.

**SEE** DBSCAN_Types.l_num_clusters
DBSCAN__Types

Go Up

IMPORTS

_versions.ML_Core.V3_2_2.ML_Core |
_versions.ML_Core.V3_2_2.ML_Core.Types |

DESCRIPTIONS

No Documentation Found

Children

1. l_stage1 : l_stage1 extends NumericField by adding a nodeID field and a fields field for the data preparation of stage 2 local clustering

2. l_stage2 : l_stage2 is the data structure for the local clustering of locDBSCAN() function

3. l_stage3 : l_stage3 is the data structure for global merging of globalMerge() function

4. l_num_clusters : l_num_clusters This record structure holds the results of functions that return statistics about the clusters formed in DBSCAN clustering, that is, it is the result structure for num_clusters and num_outliers
l_stage1 extends NumericField by adding a nodeID field and a fields field for the data preparation of stage 2 local clustering. The nodeID field records the physical cluster node to which the data point is assigned to. The fields filed allows each data point to be stored as a vector for embedded C++ computing at stage 2.

### FIELD

- **wi** ||| UNSIGNED2 — The work-item identifier for this cell.
- **id** ||| UNSIGNED8 — The record-identifier for this cell.
- **number** ||| UNSIGNED4 — The field number (i.e. featureId) of this cell.
- **value** ||| REAL8 — The numerical value of this cell.
- **nodeID** ||| UNSIGNED4 — The physical cluster node to which the data point is assigned to. It’s 0-based index by default.
- **fields** ||| SET ( REAL4 ) — The SET of feature values of each data point. It’s similar to the vector definition in C++.

SEE ML_Core.Types.NumericField.

l_stage2 is the data structure for the local clustering of locDBSCAN() function.

### FIELD

- **wi** ||| UNSIGNED2 — The work-item identifier for this cell.
- **id** ||| UNSIGNED8 — The record-identifier for this cell.
- **parentID** ||| UNSIGNED8 — the largest core points a data point belongs to.
- **nodeID** ||| UNSIGNED8 — The physical cluster node to which the data point is assigned to.
**FIELD** **fields** ||| SET ( REAL4 ) — The SET of feature values of each data point. It’s similar to the vector definition in C++.

**FIELD** **if_local** ||| BOOLEAN — TRUE if the data point is physically located at the current cluster. Otherwise FALSE.

**FIELD** **if_core** ||| BOOLEAN — TRUE if the data point is a core point. Otherwise FALSE.

---

**L_STAGE3** **l_stage3**

DBSCAN_Types \[
\]

\| **l_stage3**

l_stage3 is the data structure for global merging of globalMerge() function.

**FIELD** **wi** ||| UNSIGNED4 — The work-item identifier for this cell.

**FIELD** **id** ||| UNSIGNED4 — The record-identifier for this cell.

**FIELD** **parentID** ||| UNSIGNED4 — the largest core points a data point belongs to.

**FIELD** **nodeID** ||| UNSIGNED4 — The physical cluster node it’s located. It’s 0-based index by default.

**FIELD** **if_local** ||| BOOLEAN — TRUE if the data point is physically located at the current cluster. Otherwise FALSE.

**FIELD** **if_core** ||| BOOLEAN — TRUE if the data point is a core point. Otherwise FALSE.

**SEE** ML_Core.Types.NumericField.

---

**L_NUM_CLUSTERS** **l_num_clusters**

DBSCAN_Types \[
\]

\| **l_num_clusters**
l_num_clusters This record structure holds the results of functions that return statistics about the clusters formed in DBSCAN clustering, that is, it is the result structure for num_clusters and num_outliers. It contains the value of the statistic, per work-item

**FIELD**

<table>
<thead>
<tr>
<th>wi</th>
<th>UNSIGNED4 — The work-item identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>num</td>
<td>UNSIGNED4 — The value of the statistic (Number of clusters / outliers)</td>
</tr>
</tbody>
</table>