Data mining methods in RAVEN network

Nikolai Gagunashvili (University of Akureyri, Iceland)

nikolai@unak.is

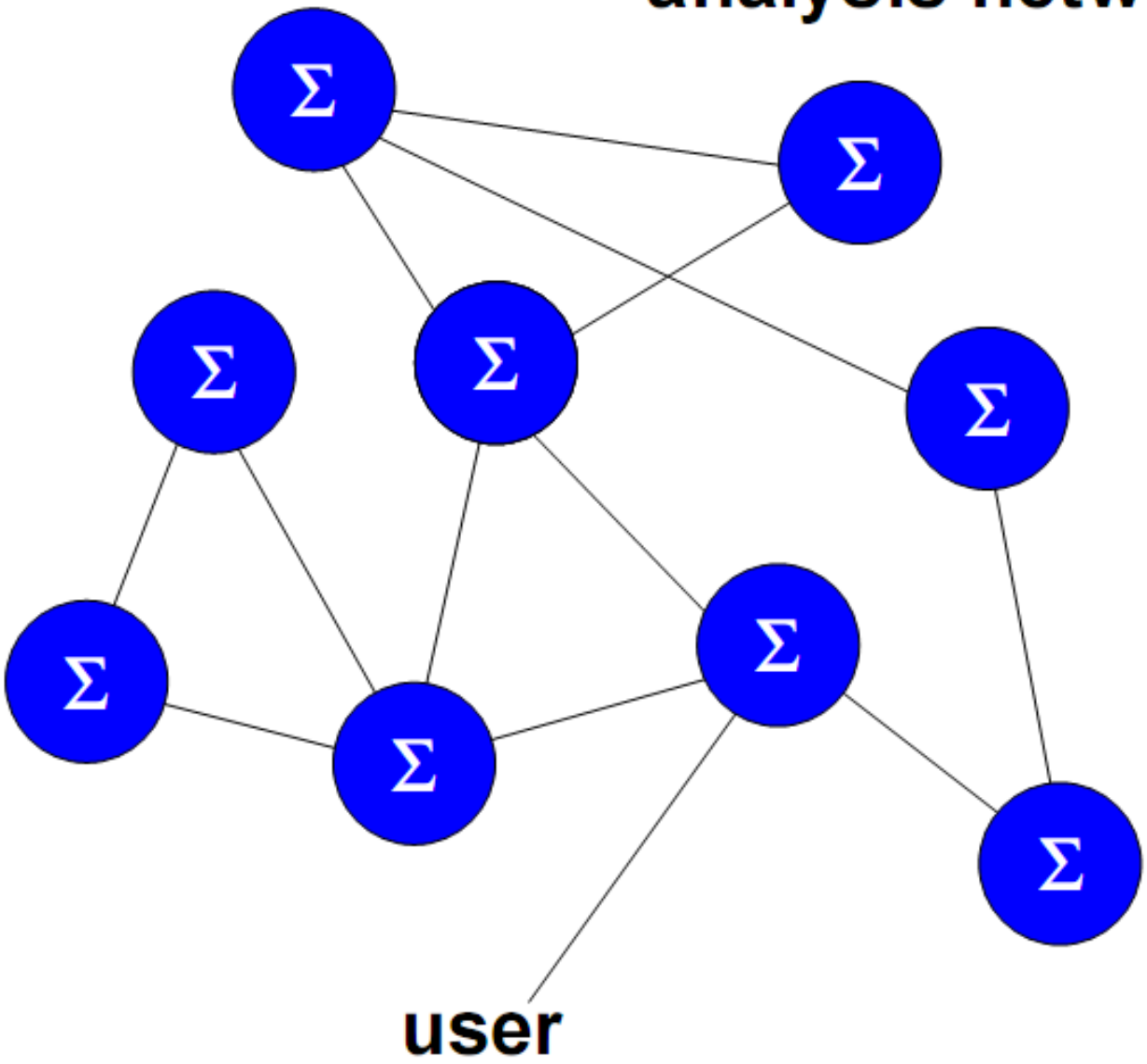
Data collected and stored at enormous speeds

Traditional techniques infeasible for raw data

Methods of data mining with visualization techniques may help scientists

- in classifying and segmenting data
- in Hypothesis Formation

analysis network

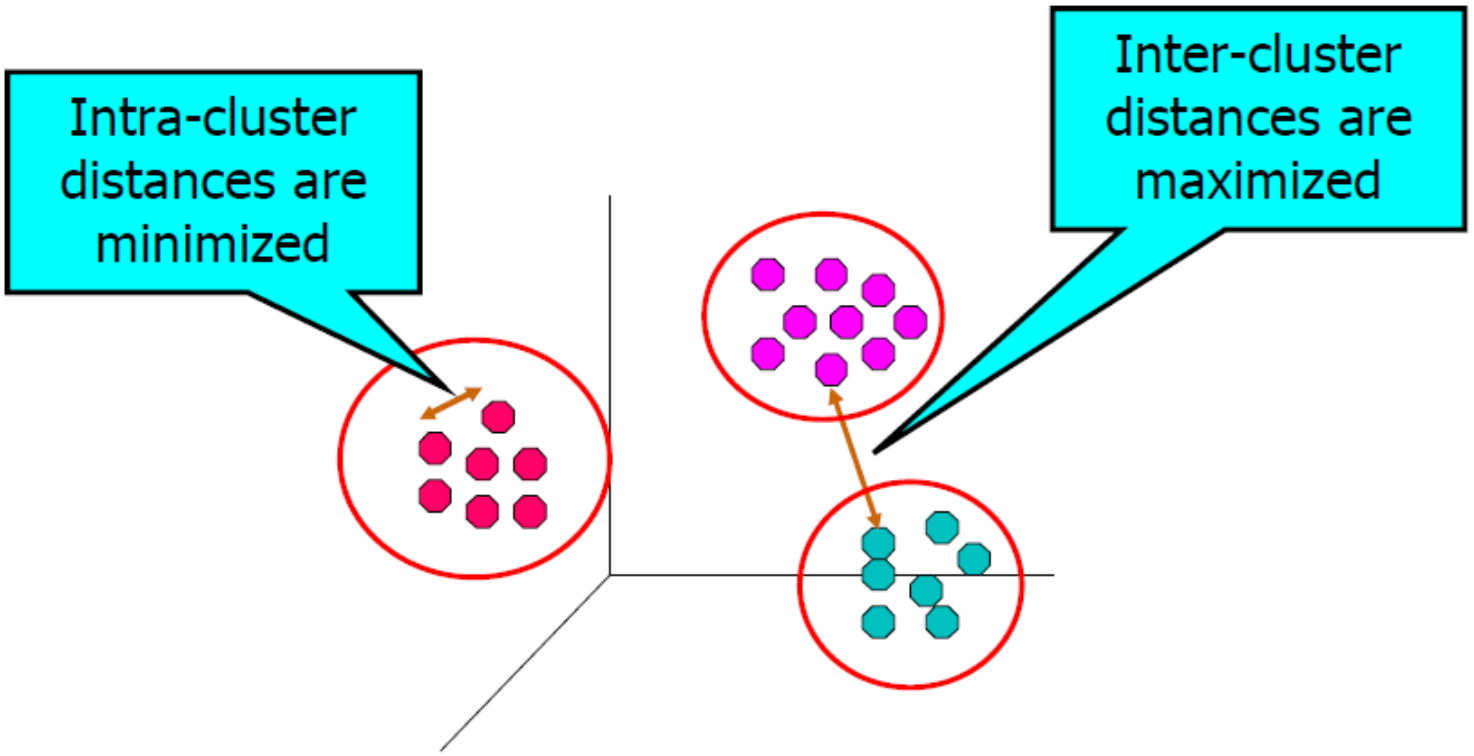


Data processing can be realized at the RAVEN network

- Unsupervised classification (cluster analysis)
- Detection anomalies (outliers)
- Supervised classification for selection rare events

What is Cluster Analysis?

Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



K-means Clustering

- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified
- The basic algorithm is very simple

Algorithm 1 Basic K-means Algorithm.

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

K-means Clustering

Algorithm 1_RAVEN Basic K-means Algorithm

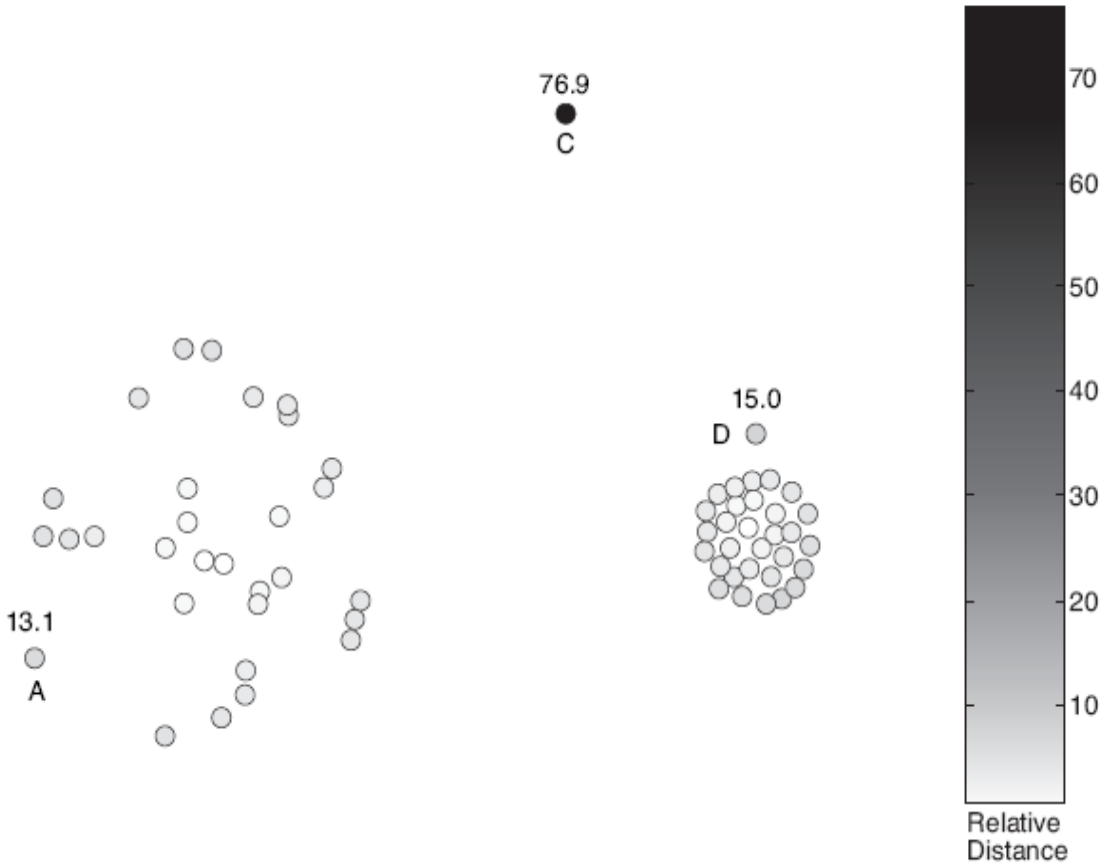
1. Select K points as the initial centroids at user node and sheare this information in network.
2. **Repeat**
3. Form K clusters by assigning all points to the closest centroid for each node.
4. Recompute the centroind of each cluster for each nodes. Output of each node are positions of centroids with number of points assigned to centroids.
5. Recalculate the positions of centroids for whole network and sheare this information in network.
6. **Until** The centroids positions for cluster don't change.

Anomaly/Outlier Detection

- What are anomalies/outliers?
  - The set of data points that are considerably different than the remainder of the data
- Variants of Anomaly/Outlier Detection Problems
  - Given a database D, find all the data points $\mathbf{x} \in D$ with anomaly scores greater than some threshold t
  - Given a database D, find all the data points $\mathbf{x} \in D$ having the top-n largest anomaly scores f($\mathbf{x}$)
  - Given a database D, containing mostly normal (but unlabeled) data points, and a test point **x**, compute the anomaly score of **x** with respect to D

Clustering-Based algorithm anomaly detection

Outlier score defined as relative distance. It is the ratio of the distance of point from the closest centroid to the median distance of all points in the cluster from the centroid.



Clustering-Based algorithm of anomaly detection

Algorithm 2 Basic clustering-based algorithm of anomaly detection

1. Find position of K centroids (Algorithm 1)
2. Find median distance for each centroid
3. Calculate the score of each pont that is
   (dist of point to nearist centroid)/(median dist. for given nearist centorid)
4. Order the scores to define outliers.

Clustering-Based algorithm of anomaly detection

Algorithm 2_RAVEN Basic clustering-based algorithm of anomaly detection

1. Find position of K centroids (Algorithm 1_RAVEN)
2. Find median distance for each centroid of network. Histogram-based method can be used for that.
3. Calculate the score of each point that is (dist of point to nearest centroid)/(median dist. for given nearest centroid)
4. Order the scores for each node.
5. Use highest scores from nodes and find points with highest score for network

Clustering-Based. Strength and Weaknesses

- **Some clustering techniques, such as K-mean, have linear or close to linear time and space complexity.**
- **Possible to find both clusters and outliers at the same time.**
- **The set of outliers produced and their score can be heavily dependent upon the number of clusters as well as the presence of outliers in the data.**
- **The quality of outliers produced by a clustering is heavily impacted by the quality of clusters produced by algorithm.**
- **The clustering algorithm needs to be chosen carefully.**

Supervised classification for selection rare events

In the analysis the LHCb focuses on very specific decay modes of B mesons which are sensitive to quantum effects caused by as yet undiscovered heavy particles ("New Physics").

Relative frequencies of B-decays 10^{-4} - 10^{-6}.

Every B-decay inside an event there are about 5-10 times that number of tracks from non-B-decays

The success of LHCb and the other LHC experiments therefore depends critically on the availability of sufficiently powerful analysis tools.

Main research must be addresses the issues of finding signals in a huge background.

The typical imbalanced problem is credit card fraud or AIDS tests. In these problems it is important to detect all rare (signal) instances if possible.

In particle selection on the other hand, performance criteria is signal/noise ratio or significance that is typical for detector devices.

The second difference is the usage of real cases in data mining in contrast to simulated training and test samples in particle physics.

Classification algorithms must therefore be robust with respect to differences in properties between simulated and real data.

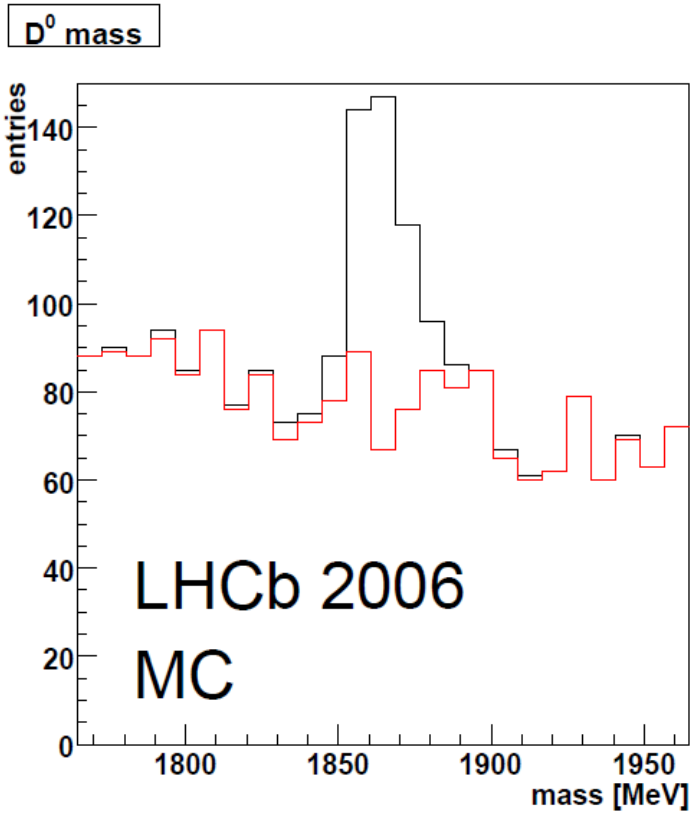
Example of selection algorithm

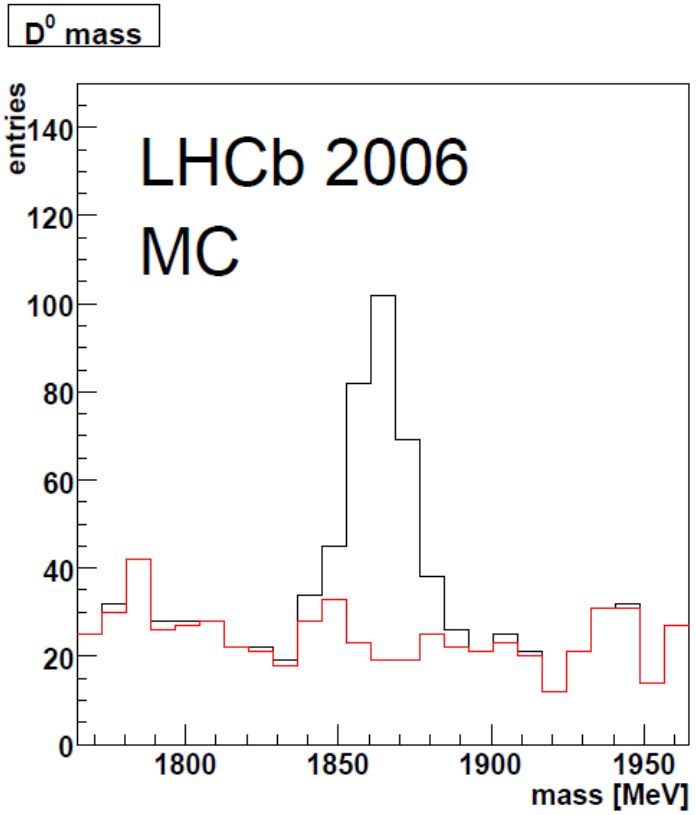
```
(IPpi >= 1.039316) and DoCA <= 0.307358) and (IP <= 0.270767) and
(IPp >= 0.800645) => class=D0

(IPpi >= 0.637403) and (DoCA <= 0.159043) and (IP <= 0.12081) and
(ptpi >= 149.2332) and (IP>= 0.003371) => class=D0

(IPpi >= 3.795743) and (DoCA <= 0.489579) and (IP <= 0.76649) and
(IP >= 0.0289) => class=D0

 => class=BG
```

IPpi, DoCA, IP, IPp, ptpi - attributes of event

D0, BG - class of event

Background/signal ~ 3000





Traditional in data analysis cut based method of selection

Selection algorithms used methods of supervised classifications

Britsch, XVII International Workshop on Deep-Inelastic Scattering and Related Subjects, 2009, Madrid

- Precision of measurements can be improved
- New method can help find particles that can not to be found by old method.
- Trigger can be organized to register only events interesting for particular physical analysis.
- New method can found application for detection complex events for example in radars, sonars, lidars technique