HOKKAIDO UNIVERSITY

Feature Selection as Reinforcement Learning Applied to Raman spectra for Cancer diagnosis

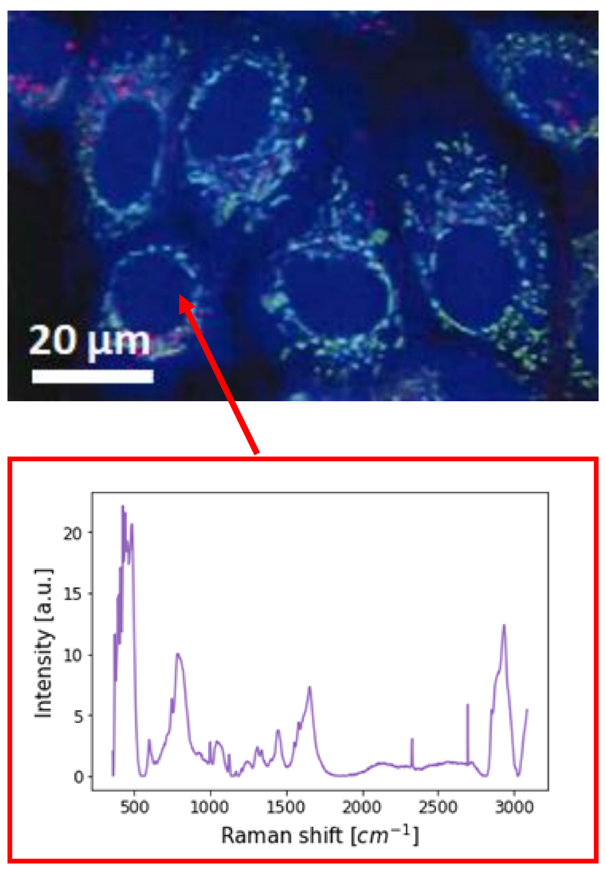
A.Pelissier, A.Nakamura, K.Tabata - June 16th 2018
Laboratory for Pattern Recognition and Machine Learning, Hokkaido University, Japan

情報認識学研究室

Abstract

There is currently an interest at finding the most relevant wavenumbers in Raman spectra from living cells for oncological applications. Information theory is used to study correlation between the wavenumbers, and feature selection methods are applied to Raman spectra to find the most informative wavenumbers for diseases diagnosis. Two different feature selection approaches based on reinforcement learning and bandit strategies are presented ; we find that 5 wavenumbers are enough to diagnosis follicular thyroid cancer with 98% accuracy.

I. Raman imaging

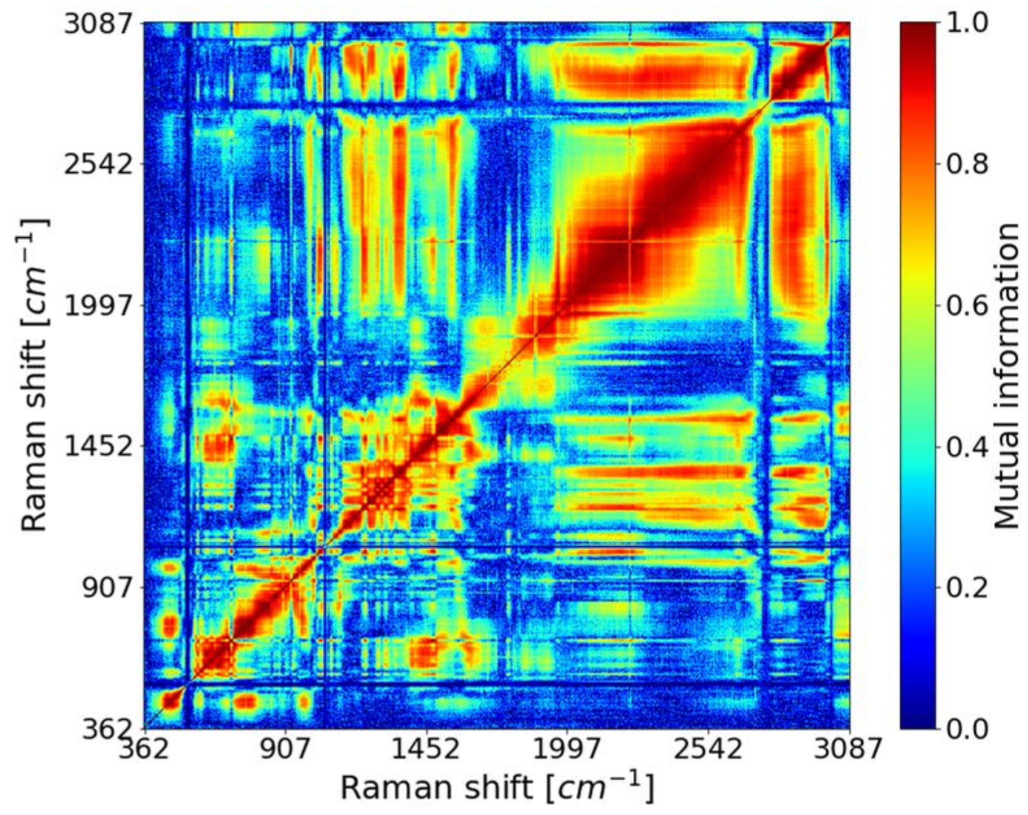


Raman measurements from living cells – Osaka group.

Raman data

- Osaka group provided 5 cancer and 4 non cancer Raman images.
- A Raman image contains 400 x 250 spectra at different positions.
- Each spectrum contains 840 wavenumbers.
- Each peak indicate the presence of a specific molecule, notably present in lipids and proteins.
- It is not possible to diagnosis cancer by eye.

II. Wavenumbers mutual information



Normalized mutual information between Wavenumbers in the Raman spectra

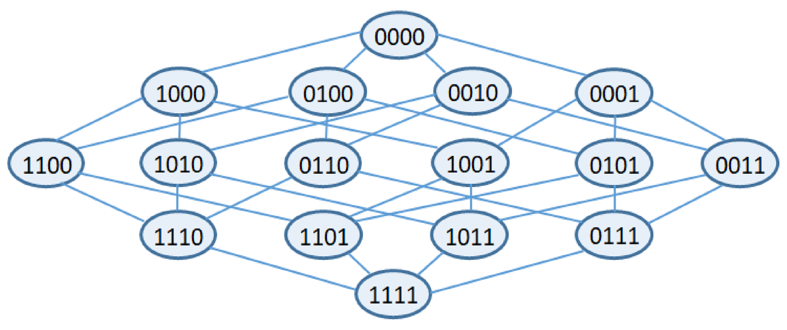
High redundancy

- The mutual information between wavenumbers is calculated from 2500 Raman spectra.
- High mutual information within many wavenumbers is observed.
- The wavenumbers seem to be splitted into different information clusters.
- Some wavenumbers that are far away in the spectra are highly correlated.

III. Feature Selection as Reinforcement Learning

Feature lattice

- Each node in the feature lattice corresponds to a unique feature subset.
- The lattice corresponding to a feature set of size f contains 2^f nodes.
- A node at depth d in the lattice has d parents and $f-d$ children .



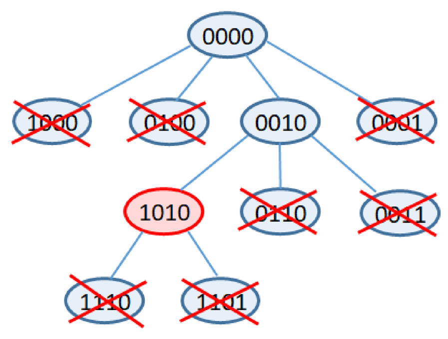
Example of a feature lattice with a feature set of cardinal $f = 4$, containing $2^4 = 16$ nodes.

Fast feature set evaluation

- Feature subsets are evaluated by a k Nearest Neighbor classifier trained with the selected features.
- To reduce the computational cost when dealing with large dataset, a small subsample $\mathcal{V}$ of the original set is computed.
- The score of the feature subset is taken as the Area under the ROC curve of the kNN predictions on $\mathcal{V}$.

Greedy

- Start from the empty feature set.
- Repeatedly add the best additional feature to the set.
- Stop when no additional feature further improves the set evaluation.



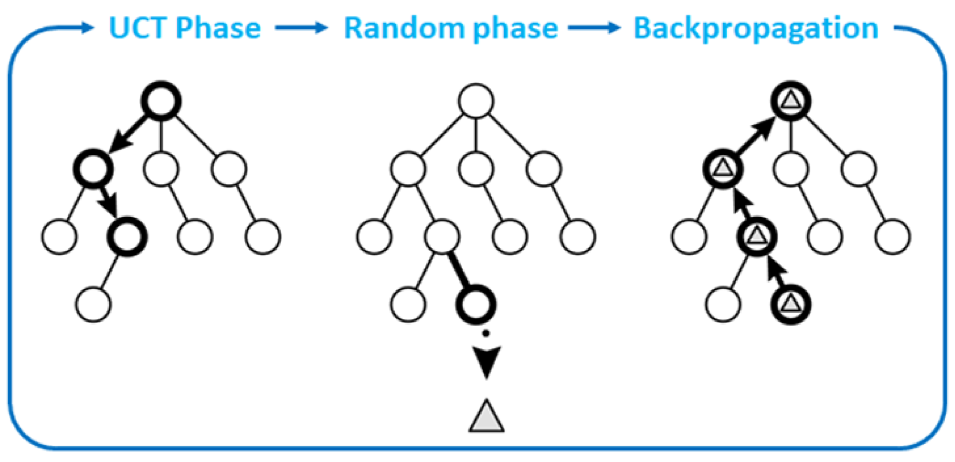
Example of a Greedy search

FUSE

Iterates N times with the following process:

- **Selection**: Starting at the root node, the UCT selection policy is recursively applied to descend through the tree.
- **Simulation** is run according to the random policy to produce an outcome.
- **Backpropagation** through the selected nodes is performed to update their statistics.

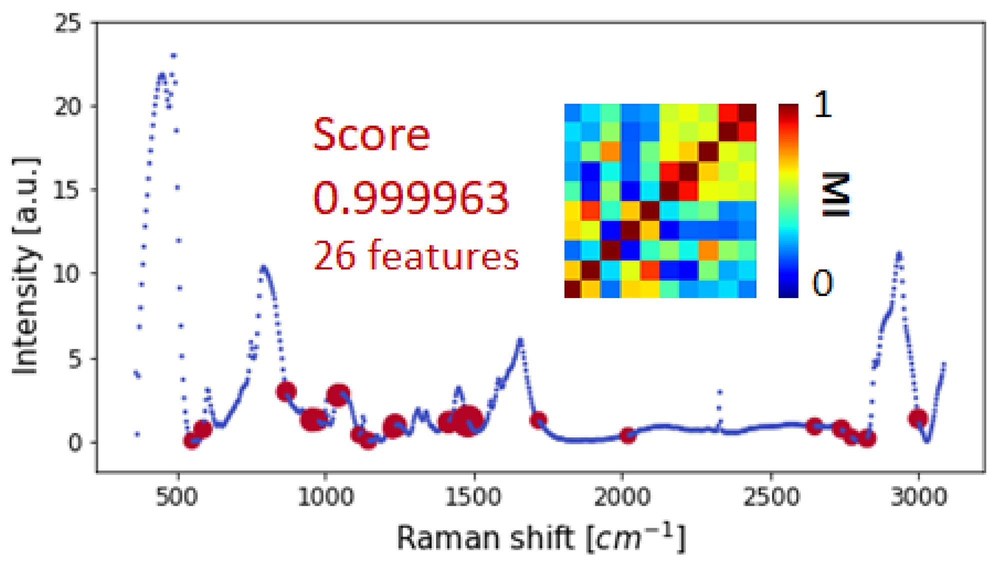
The best node is at the end of the most visited path.



One iteration of the FUSE algorithm.

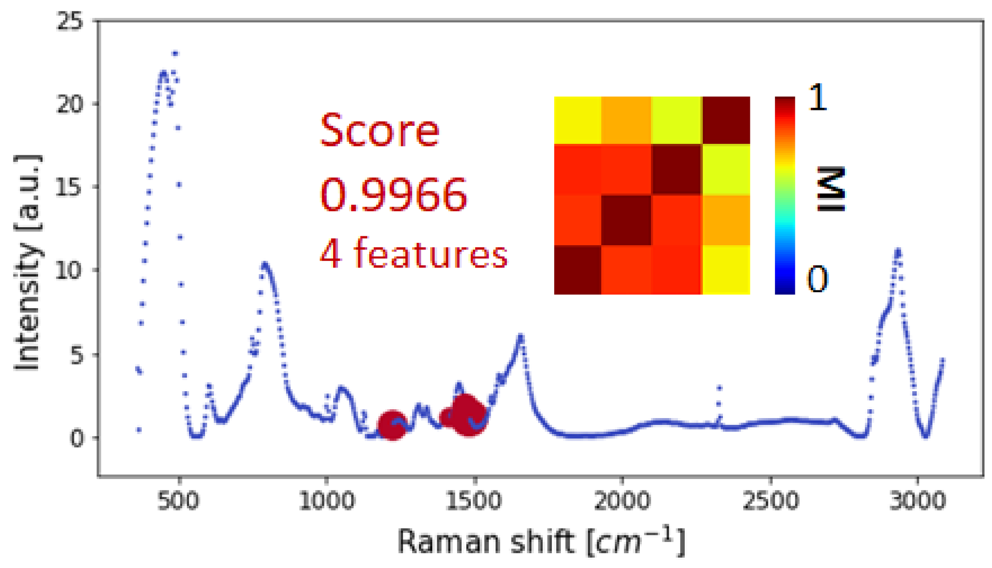
IV. Wavenumber selection on Raman spectra for follicular thyroid cancer diagnosis

Greedy results



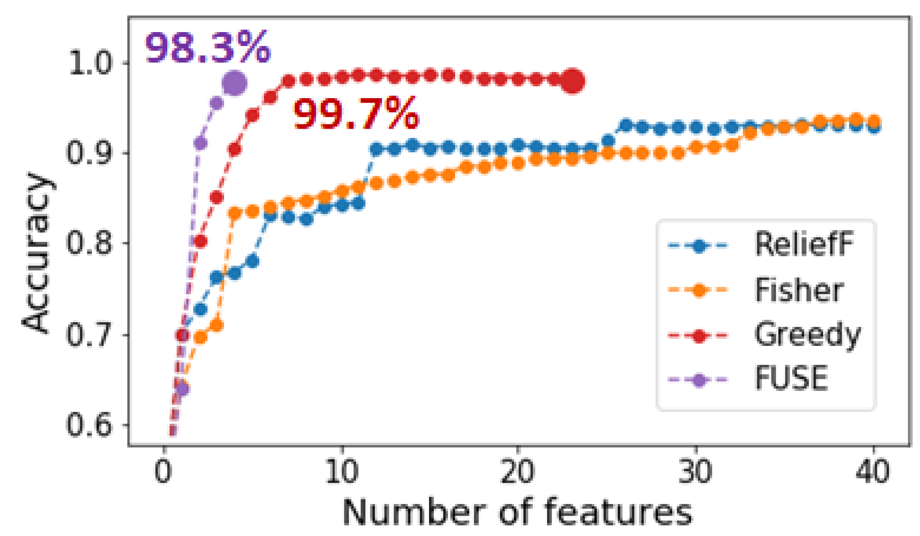
Results of the Greedy algorithm on the Raman spectra.

FUSE results



Results of the FUSE algorithm on the Raman spectra.

Algorithms performances



Accuracy obtained with a 5 Nearest Neighbors classifier trained with different feature set.

- Both FUSE and Greedy approach performed much better than major Filtering methods (ReliefF and Fisher).
- FUSE is able to find a better *combo* of features than Greedy, but fails to select additional features when the feature set evaluation becomes close to one.
- The mutual information of features selected by FUSE and Greedy is surprisingly high, indicating that each redundant features brings their own contribution for cancer diagnosis.

5 wavenumbers from the original 840 initially contained in the Raman spectra are enough to predict cancer with 98% accuracy.