Priority Based Functional Group Identification Using Machine Learning

Rushikesh Nalla, Rajdeep Pinge,
Manish Narwaria, Bhaskar Chaudhury

Introduction

An organic molecule contains different functional groups. There exists a dominant functional group in the molecule which determines majority of its properties and its reactivity. FT-IR is a spectroscopic method which gives spectra of molecules based on their functional groups. The aim is to identify the patterns in the spectra to determine the dominant functional group.

Visual Idenification

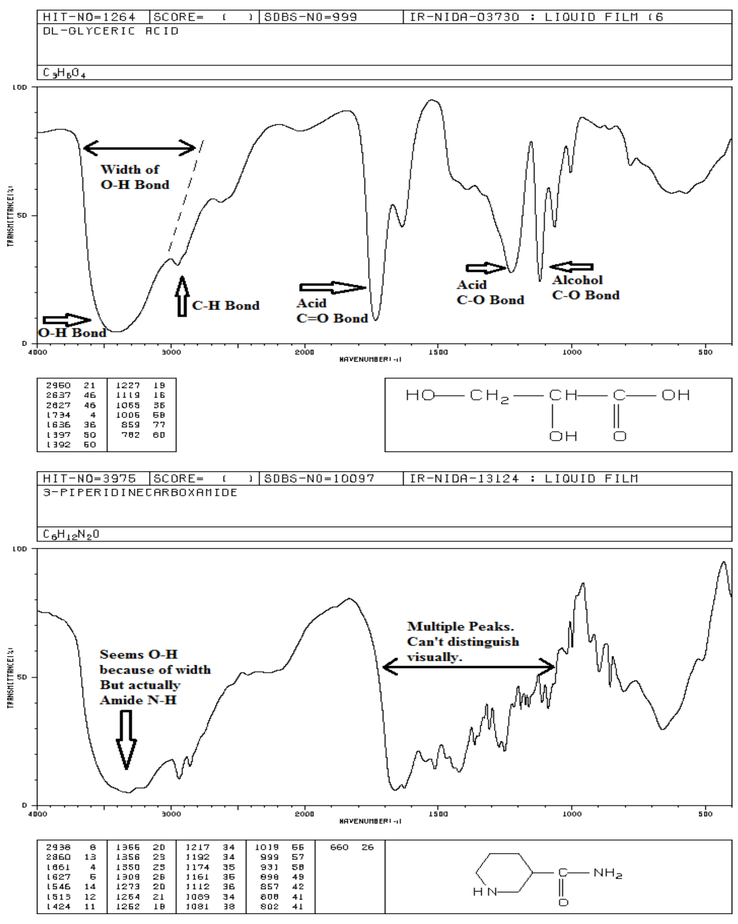

Fig. 1 Problems in Visual Identification.

Data Collection

1341 spectrum samples belonging to 14 functional groups collected from SDBS database.
Challenges:

1. Data in image form: Process pixel by pixel to quantify data
2. Two different scales on x-axis

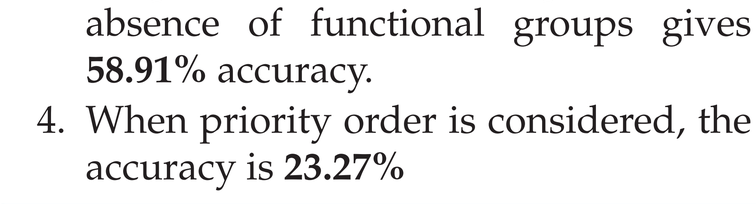

Fig. 2 (1) Transmittance Level, (2) Width of Peak, (3) Number of Peaks in the given Range (4) Sum of Widths of Peaks.

Rule Based Approach

1. Similar to visual identification process.
2. Error-prone if spectra are complicated with no distinguishing features.
3. Simple check about the presence or absence of functional groups gives **58.91%** accuracy.
4. When priority order is considered, the accuracy is **23.27%**

Data Driven Approaches

Chemoinformatics $\begin{cases} Machine\ Learning \\ Spectroscopy \\ Chemistry \end{cases}$

ML Based Approaches

Instead of fixed rules, allow machine learning algorithm to extract the distinct patterns based on data.

1. **Single-Label Intermediate Approach (SLIA):** 4 features for each of 23 bond ranges used in rule based approach. Single Class Label.
2. **Multi-Label Multi-class Approach (MLMC):** Extract 250 features at regular intervals from spectroscopic sample. Multiple class labels. 14-label vector with one label for each class.
3. **Single Label Multi-class Approach (SLMC):** Extract 250 features at regular intervals from spectroscopic sample. Single class label representing the dominant functional group.

Results

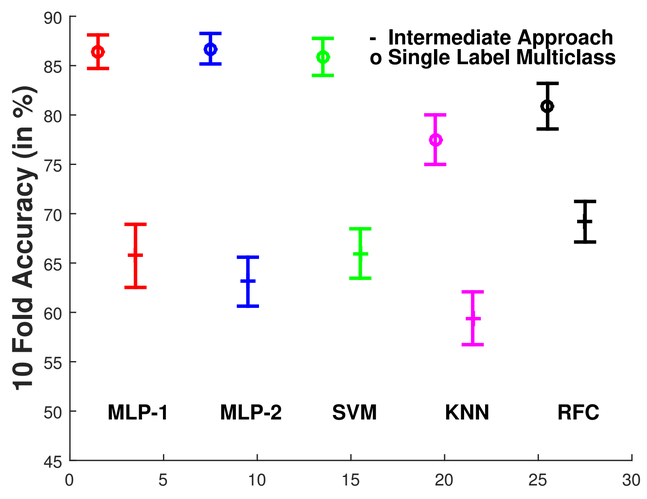

Fig. 3 Comparison of SLIA and SLMC Approaches

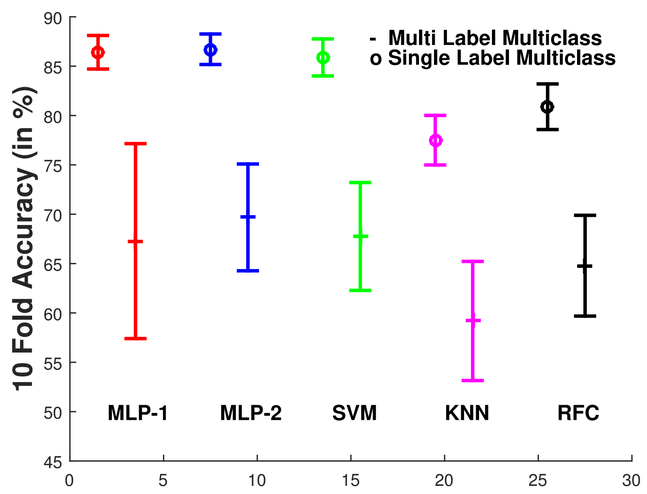

Fig. 4 Comparison of MLMC and SLMC Approaches

Contact

Group in Computational Science and HPC, Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar - 382 007, Gujarat (India)
Tel.: (+91) 079-3051-0590
Email: bhaskar_chaudhury@daiict.ac.in
Website URL: http://cshpc.daiict.ac.in/

Statistical Analysis

1. **Class Permutation Test -**
   Null Hypothesis: Class structure is not exploited by the algorithm (no dependence between class label and features)
   Result: p value is less than 0.01 for all algorithms.
2. **Feature Permutation Test -**
   Null Hypothesis: Feature dependence is not exploited by the algorithm (no dependence among the features)
   Result: p value is 1 for all algorithms except KNN whose p value is 0.01.

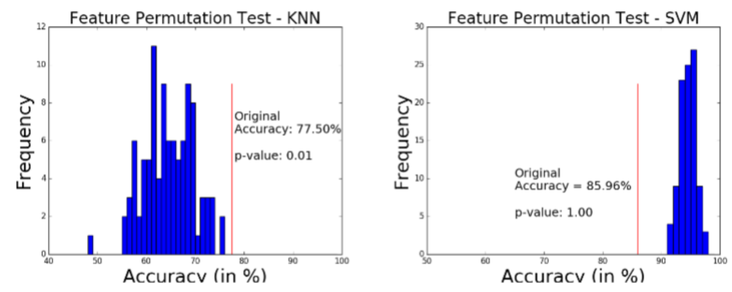

Fig. 5 Results of Feature Permutation Test

Conclusion

- Goal is to assign a single label which is the most dominant functional group present in the sample.
- Rule based approach gives a poor accuracy of **23.27%** in priority based classification.
- SLIA gives the least accuracy **(60-65%)** among the three ML based approaches.
- The accuracy improves slightly **(65-70%)** in MLMC approach and is maximum **(80-85%)** for SLMC approach.
- **Our proposed SLMC approach works better than the previously used MLMC approach.**

Future Work

- More insights into high p-values of feature permutation test.
- More experiments with white box approaches to get better insight into groups which perform poorly and understand why they do so.

Acknowledgement

The authors express their sincere gratitude to National Institute of Advanced Industrial Science and Technology, Japan for open source SDBS database of FT-IR spectra.

References

1. Baker, Using Fourier transform IR spectroscopy to analyze biological materials, Nature Protocols, 2015.
2. Robb, E. W. and Munk, M. E. A, Neural Network Approach to Infrared Spectrum Interpretation, 1990.
3. Ojala, Permutation Tests for Studying Classifier Performance, J. Mach. Learn. Res., 2010.