Total Molecular Structure Analysis from Fourier transformation infrared spectroscopy (FT-IR) using Machine Learning Approach

Conference Paper · September 2020

CITATIONS
0

READS
24

2 authors, including:

Anye Shi
University of California, San Diego
5 PUBLICATIONS **108** CITATIONS

SEE PROFILE

Total Molecular Structure Analysis from Fourier transformation infrared spectroscopy (FT-IR) using Machine Learning Approach

Anye Shi[1 2]
1 Electrical and Computer Engineering Department
2 Department of Chemistry and Biochemistry
University of California, San Diego
anshi@ucsd.edu

Shuai Zhang[1 3]
1 Electrical and Computer Engineering Department
3 Mechanical and Aerospace Engineering Department
University of California, San Diego
shz134@eng.ucsd.edu

Abstract

Multi-label classification analysis of molecules' structure based on FT-IR spectrum is reported in the paper. Approximately 600 different organic molecules with C, H, O elements were well collected from National Institute of Standards Technology (NIST) database with their molecular structures, CAS numbers and standard infrared (IR) spectrum. Nine functional groups were found in these molecules and data were labeled based on these functional groups in the molecules. Before multi-label classification, decision tree, support vectors machine (SVM) and multi-layer perceptron (MLP) algorithms were implemented for one-versus-rest binary classification and the results were compared. Multi-label classification with both problem transformation method (binary relevance) and adaptive algorithm methods (Multi-label k Nearest Neighbours (MLkNN)) were then introduced to identify all functional groups in each molecule by their IR spectrum. Lastly, 50 Alkane molecules with different molecular length were chosen to simulate their molecular structure and compute their IR spectrum to study what happened for IR spectrum when molecular length grows. A general principle of relation between IR spectrum peak shift and number of bond length was obtained by using multiple regression methods which could be used for peak position correction in future work to enhance the classification accuracy.

1. Introduction

Infrared (IR) spectroscopy is among the most powerful methods for determining the structure of all kinds of organic compounds due to their unique vibration transition state structure on atomic level. Typically, different types of chemical structure will exhibit characteristic signal position in IR spectrum. This has led to its broad use today in many fields including for instance materials and pharmaceutical chemistry. In the latter the determination of structure and packing is essential to elaborate structure–property relations for formulations in the drug development process.

However, as molecular structure goes to larger and more complicated, it can be difficult to confidently assign peaks to their geometry structure from IR spectrum. Because we currently assign IR peak to their structure based on some empirical guidelines manually, which are not always reliable. It is crucial to build up a more accurate way to analyze and assign peaks to structures automatically. In addition, there are some peak regions overlapping for different types of chemical structures. It would be hard to differentiate the origin of some certain peaks if we have two chemical structures with similar IR-response signal ranges.As you can see in Fig. 1 , there are several reference absorption peaks overlapping in the wave number range from 3000 to 3500 cm^{-1}. What's more, IR peaks could also shift if we perform the characterization in different reaction conditions.

Another shortcoming of IR spectrum analysis is the failure of determining the size of molecules. Since IR spectrum is a typical qualitative characterization and we

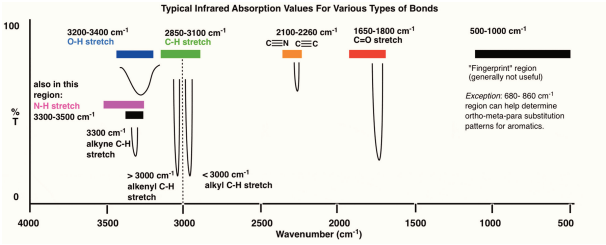

Figure 1. Approximate FT-IR peak ranges for different chemical structures

could only get the structure information from IR spectrum. We can't get any quantitative information (molar mass, molecular size) from IR spectrum. This intrinsic drawback makes IR analysis only be a supporting characterization of organic compounds. We still need to perform other quantitative characterization to fully understand the compounds structure.

Machine learning (ML) is emerging as a new tool in many areas of chemical and physical science, and potentially provides a method to bridge the gap between the need for high accuracy calculations and limited computational power. Various machine learning methods have been applied to IR spectrum analysis studies over the years. Ellis *et al.* applied a multiple linear regression method to classify different types of muscle foods (beef, lamb, pork, chicken, turkey) based on the corresponding IR spectra [3]. Howley *et al.* classified chemical samples containing acetaminophen using SVMs and k-means with a PCA-reduced Raman spectra feature sets [4]. Zou *et al.* trained a SVM model on near-infrared (NIR) spectra for the identification of oxytetracycline powder [19]. Despite the booming interests in applying machine learning classification methods to spectral data, most of literature to date seeks to either classify for a highly specific compound or focus on molecules with very simple structures, which makes the methods limited to specific cases. What's more, since more than one function groups exist for most of the molecules, single label classification are not able to meet the needs for material identification especially for the molecules with complicated structures. In this paper, we tried several different algorithms to classify the molecules with multiple labels. The methods can be used to any molecules with function groups mentioned in the paper. The peak shifting problem caused by increasing molecular weight is also studied with several regression techniques which would be helpful for building a classification model with better accuracy in the future.

2. Methods

2.1. Data preparation

The standard IR spectrum and basic information of organic compounds were scraped with Beautiful Soup from the NIST Chemistry Webbook[12]. Since there are too many functional groups with different types and some of them are rare with very few samples to train. In this study, we only focused on the molecules with C, H and O elements and found 9 different functional groups (shown in Fig. 2) in these molecules. After filtering the scraped information based on the elements, effective information for 572 molecules were stored.

The values on each peak position were used as features for training machine learning algorithms. One problem from the original data from spectrum is that the position of the points and range of wavelength are random based on the researchers' personal preference. To unify the dimension of the training data, we selected the range from 450 cm^{-1} to 4000 cm^{-1} and calculated the intensity of spectrum at integer positions. Two nearest points around the integer were searched in the data set and the intensity on the integer position was calculated by weighted average of intensity of these two points. Weights were determined by the distances. To be mentioned, regression algorithms for peak shifting study requires accurate peak positions. Thus, instead of using all the data in the spectrum, the peaks need to be labeled to extract peak position information for regression tasks.

The rules of naming molecules, especially for those with complex structures, are complicated and also vary with different types of compounds. It is hard, if possible, to label all the functional group purely from the information from the names themselves. Since the exhaustive labeling of functional groups for molecules are not done before, we manually labeled all the functional groups.

Except getting spectrum directly from database, we also used another accurate method to calculate IR spectrum, which is plane wave density functional theory (DFT) methods. DFT allows people to calculate the spectrum based on the structure of molecules. This method was used to generate spectrum for molecules with incomplete spectrum information from the database. However the computation expense is prohibitive.

2.2. Binary classification

Most traditional learning algorithms are developed for single-label classification problems. There are also literature transform the multi-label problem into multiple single-label problems, so that the existing single-label algorithms can be used as base classifiers in the multi-label tasks. For binary classification, decision

tree[8], SVM [16] and MLP[10] were used and compared. Grid search technique was implemented to find the best combination of hyper-parameters. To prevent over-fitting, we used 5 fold cross validation in all the training processes. The average F1 scores for the test data with different train data and test data combination were calculated to evaluate the models.

2.3. Multi-label classification

Scikit-multilearn package [14] was used for multi-label classification. A frequently used method for multi-label classification is to do multiple binary classification independently and combine the results together. The algorithms built based on this principle are called problem transformation methods [11][2], since these methods transform multi-label problem to binary classification problem. Another adaptive algorithm method was also implemented in the paper to compare with problem transformation method. Adaptive algorithms are specially designed for multi-label learning tasks [9][1] and multi-label k-nearest neighbor [17] was chosen to classify functional groups based on the spectrum information. Hamming-Loss and exact-match ratio were the two terms to evaluate the accuracy of the models. Hamming-Loss is the fraction of labels that are incorrectly predicted, i.e., the fraction of the wrong labels to the total number of labels. The requirement for the exact-match ratio is much higher, which need all the labels to be correct in the prediction.

2.4. 4. Peak position prediction with increasing repeating structures

Due to the intrinsic properties of organic molecules, not all the peaks tend to shift when the molecular length enlarged. So we select 8 peaks which are sensitive to molecular size. 49 Alkanes with carbon number from 2 to 50 were taken as research object. IR spectrum were obtained from QM simulation. The reason why we took the simulation results for further calculation is that: (a) Solvent effect and background noisy signal will introduce much error for peak assignment. (b)QM simulation is also called first-principle simulation with highest accuracy to describe electronic and vibrational structure for single molecules.

Different regression methods were used to deal with this problem. Typically, we have different types of Gaussian Process Regression models, SVM regression, linear regression models, tree-ensemble methods and multi-parameter regression. Root Mean Square Error (RMSE) and R-squared value are used to describe the model performance.

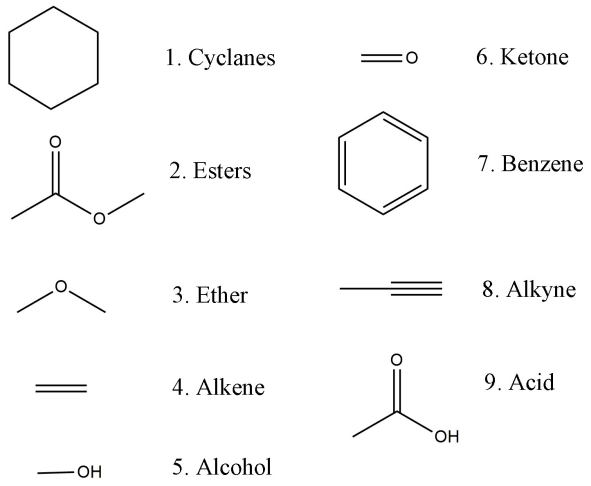

Figure 2. Classification principle for compounds list obtained from NIST database.

2.5. DFT based computational simulation details

Monte Carlo conformational molecular dynamic (MD) simulation are studied using *Vienna Ab initio Simulation Package (VASP)* with D3 computation basis Density functional theory (DFT) based quantum mechanics simulation are employed using a custom-written basin hopping (BH) algorithm to search the cluster potential energy landscape and identify candidate structures for treatment at the B3LYP/6-31g level of theoretical accuracy. The geometry optimization and frequency calculation are performed in *Gauss09*.

3. Results and discussion

3.1. Binary classification

3.1.1 Decision tree

The decision tree is a supervised machine learning algorithm, which regards the whole data as root and split the data based on rules. Theta Automatic Interaction Detection (THAID) [7] is the first paper in which extends the idea of tree to classification. For continuous data like we had in FT-IR spectrum, gini index was used as criterion to calculate values for every attribute. The values were sorted, and the attribute with the highest gini value was regarded as the root and the separation rules (where to split) were also made sure based on the values. Top down decision tree was first introduced by Quinlan *et. al* [8]. The advantage of decision tree is that the process of learning is similar to decision making process of human, which makes the algorithm more explainable comparing to many other algorithms.

Max depth of the tree is one of the most important hyper-parameter in the model. The depth of the tree de-

Functional group	Cyclanes	Esters	Ether	Alkene	Alcohol	Ketone	Benzene	Alkyne	Acid
Optimized F1 score of DT [%]	59	85	69	72	84	75	82	76	76
Optimized tree depth	6	2	3	9	2	1	1	4	2

Table 1. Optimized results of decision tree and corresponding tree depth

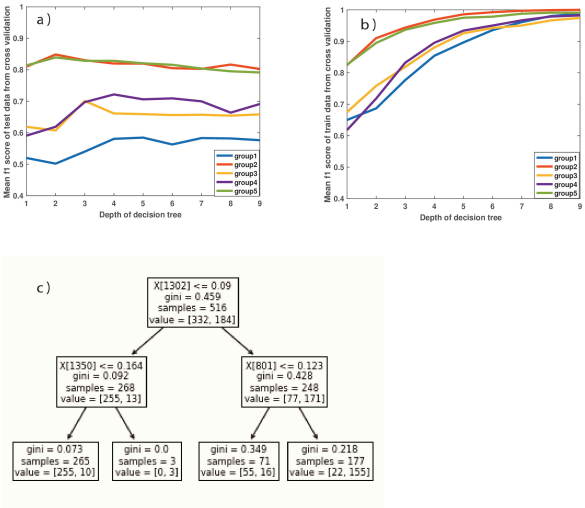

Figure 3. (a,b) Test and train data f1 score with increasing decision tree depth. (c)Decision tree structure for 'Esters' functional group

cides how many times would the splitting process to be implemented, the more the splitting, the more complicated the model is. As can be seen from the Fig.3(a) and (b): as the depth of decision tree increases, the accuracy of training data keep increasing (for all functional groups 1-5). It is obvious that depth of tree enlarges the complexity the model and tends to overfit the training data set. However the test data score is not monotonically increasing. That's because the test data scores were got by taking averages of test data accuracy with different combinations of train and test data, which prevents the overfitting from happening. From table 1, functional group 'Cyclanes' and 'Alkene' require more tree depth but the accuracy are two of the lowest ones. This gives us the insight that the classification tasks for these two groups might be more complicated than others. We also plotted the tree structures for all the classification models with optimized parameters. An example of plotted tree structure for 'Esters' is in Fig.3 (c). What's interesting is that the features (wave number) that differentiate the samples better are out of the range of their own spectrum peak range. In another word, decision tree model is classifying by determining which wave number position with high intensity would be impossible for one functional group. This is because though there are suggested ranges for different structures based on experiment results, the peak themselves are too narrow. Also, some of the ranges overlap with each other and position of peaks shift with different surrounding structures. It is almost impossible to use intensity at a single position to classify the molecular structures. For 'Esters', C-O bonding peaks are in range of 1000-1300 cm^{-1}, while the first features used to split the data are position at 1302, 1350 and 801 cm^{-1}. These positions are out its C-O spectra range but help separate 'Esters' from other structures with peaks close to the range.

3.1.2 SVM

Support vector machine (SVM)[16] is recognized as one of the most powerful kernel-based tool for classification, which has being applied to many fields successfully[13]. SVM has been proven to be a powerful methodology to perform nonlinear classification, multivariate function estimation, and nonlinear regression [15]. In this method kernel maps the data into a higher dimensional input space and constructs an optimal separating hyperplane in this space. Recently, SVM technique has been employed to an extensive application for discrimination. Zhao *et al.* [18] utilized IR spectroscopy combined with SVM to identify green, black, and oolong teas. Langeron *et al.* [5] used SVM to classify IR spectra of tissue samples. One of the typical advantages of SVM, when compared with other methods, is that there are very few parameters to tune or select a priori. We first initialize our experiment by using parameters with C=100, gamma=0.10 and max iteration=10000). The optimized parameters for SVM are determined to be C=200, gamma=0.02 and max iteration=10000 after running Bayesian Optimization. Three different modes of SVM classification (linear, rbf and sigmoid SVM) were chosen for comparison.

The most straightforward results are shown in Table 2. The SVM binary performance could be evaluated roughly from accuracy values. All the structures have accuracy values that are higher than 85 percent, indicating good performance for overall models. The classification results of Structure 1, 4, 5 are less satisfying. It could be ascribed to huge spectrum overlap between these three structures. Cyclanes (Structure 1) has strong absorption from 2900-3040 cm^{-1} due to multiple C-H stretching vibration structure. Alkenes (Structure 4) exhibit structure similarity with Cyclanes, which is sensitive to the IR range from 3000-3100 cm^{-1}. And al-

cohol (Structure 5) have a broad absorption peak from 2700-3600 cm^{-1} due to strong peak splitting and degeneracy. Highly overlapped spectrum makes it harder to differentiate each peak and assign peaks to their proper structures. The complexity of the classification for each functional group can also be evaluated with number of support vectors and group 4 and 5 has largest number of support vectors based on the result from Table 2. This indicates that more complicated model is needed to classify group 4 and 5 from the rest of the structures. To be mentioned, the model for group 4 is also complicated based on results from decision tree algorithm. We also plotted accuracy value obtained from three different SVM models in Figure 4 (a1). Linear SVM and rbf SVM model work better for all the samples compared with sigmoid SVM. It indicates that sigmoid function might not be a proper activation function to classify IR spectrum into different groups. It is interesting to find that linear SVM model has better activity when we were performing classification for carbon-riched structure while rbf SVM shows good performance to deal with oxygen-related structure. To better exhibit the accuracy gap, we then plotted Figure (a2) with y axis as accuracy of (rbf method - linear method). A positive value shows rbf SVM are more efficient while negative value indicates linear SVM is more suitable to conduct. To dig out the reason beneath the phenomena, we exhibit the F1 score for linear SVM and rbf SVM in Table 3. F1 score for group 1 (g1) generally tells us that we correctly put Structure i in Class i (we correctly assign structure to their proper class). And on the contrary, g-1 typically shows we didn't put structure into class it doesn't belong to (We didn't classify structure in wrong category.) Structure 1, 4, 7, 8 are what we called carbon-rich structure. And we could notice a much better F1 score for linear SVM model of g1 group when it refers to Structure 1, 4, 7. It indicates that we are more confident to assign carbon-rich structure to their specific class correctly by using linear SVM model.

3.1.3 MLP

It has been noticed that supervised multivariate classification strategies, for example by multilayer perceptron (MLP) neural network (NN) models with supervised learning are the techniques of choice for the development of effective and robust classifiers for IR-based classification of tissue structures [10]. These supervised techniques can be efficiently optimised by pre-selecting the appropriate spectral features from the spectral data [6]. Neural networks can be "trained" to solve problems that are difficult to solve by conventional computer algorithms. Training refers to an adjustment of the connection weights, based on a number of training examples that consist of specified inputs and corresponding target outputs. Training is an incremental process where after each presentation of a training example, the weights were adjusted to reduce the discrepancy between the network and the target output Parameter optimization of MLP classification were employed using 'GridSearchCV'. The final parameters were determined to be hidden layer size (50,50,50) with Relu being activation function and stochastic gradient descent (sgd) being the solver. Alpha value was set to be 0.001 with learning rate interval as 1. The learning curves for all 9 classes of structures are shown in Figure 4 (b1,b2). Most of the samples converged within 100 iterations but the loss curves for group 1, 4, 5, 7 are not as smooth as others. We then calculated the score for training data and test data with their converged iteration times. All the results are shown in Table 4. We have similar results for MLP compared with SVM method. Structure 1, 4, 5, 7 can't be specified well using MLP method. Poor classifier performance for structure 1, 4, 5 are explained in SVM part. Bad results for structure 7 could be attributed to insufficient training. We could notice that the training score for structure 7 is pretty low compared with the other 8 classes. Low training score generally indicate that we didn't get enough distinct features to specify samples efficiently. In order to avoid this phenomena for next time, we should add more samples for class 7 with more distinct structure.

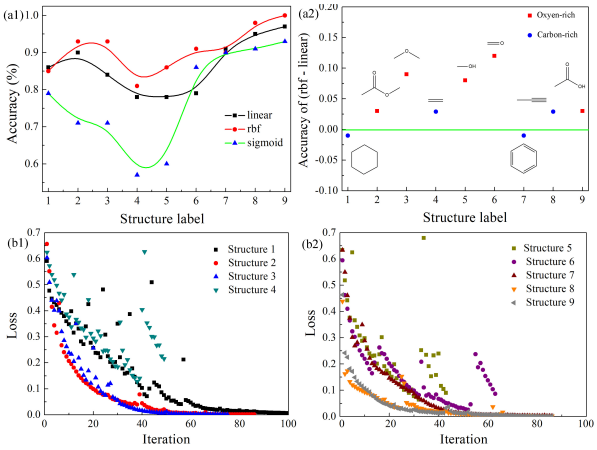

Figure 4. (a1,a2) Accuracy of optimized SVM classification on each functional group (b1,b2) MLP loss change with increasing iteration numbers

3.2. Multi-label classification

Since SVM had a good performance on the binary classification, we used SVM as independent base classifier for binary relevance algorithm. To study the relationship between single classifier and combined re-

Functional group	1Cyclanes	2Esters	3Ether	4Alkene	5Alcohol	6Ketone	7Benzene	8Alkyne	9Acid
Accuracy of SVM [%]	86	93	93	81	86	91	91	98	97
Number of SV	188	300	163	453	431	96	8	54	4

Table 2. Summarized results of support vector machine (SVM).Highest accuracy is selected among three values obtained from linear SVM, rbf SVM and sigmoid SVM. The number of support vectors (SV) are exhibited.

Functional group	1Cyclanes	2Esters	3Ether	4Alkene	5Alcohol	6Ketone	7Benzene	8Alkyne	9Acid
F1 for linear SVM of g1	0.62	0.85	0.67	0.63	0.55	0.5	0.86	0.64	0.75
F1 for rbf SVM of g1	0.43	0.90	0.86	0.56	0.75	0.8	0.84	0.67	1.00
F1 for linear SVM of g-1	0.90	0.92	0.90	0.84	0.85	0.87	0.94	0.97	0.98
F1 for rbf SVM of g-1	0.92	0.95	0.95	0.88	0.90	0.98	0.92	0.99	1.00

Table 3. F1 score of support vector machine (SVM) classification.We label group 1 (g1) for our as-studied structure and group -1 (g-1) as other structures.

sults, we plotted in Fig. 5 (a) the mean accuracy for 9 functional groups, the accuracy for 'alkene' functional group and exact-match accuracy for binary relevance classification results with different gamma value (C value remains the same as that in optimized SVM model trained above). The reason why we included accuracy for 'alkene' is that it is the functional group that is most sensitive to the change of the gamma among all the groups. While the mean accuracy value of 9 functional groups slightly decrease with increasing gamma value, the change of exact-match accuracy decreases much faster. The trend of exact-match accuracy closer to that for 'alkene' accuracy changes. It makes sense because for exact-match accuracy even one single incorrect prediction would be regarded as failure so the target that is more sensitive to the change would affect the accuracy to a greater extent. The Hamming loss for the model after optimization is 0.1 and exact-match accuracy is 0.49.

One disadvantage of binary relevance approach is that the method doesn't take relationship among labels into consideration at all. The classification task on each label are independent to each other. While adaptive algorithms are designed to fix the problem. Multi-label k nearest neighbor method [17], in the category of adaptive algorithm, is used in this paper to classify multiple functional groups in the molecules. Two most important parameters to be tuned in the MLkNN are k (number of neighbors of each input instance to take into account) and smoothing parameter. Figure 5 (b) shows the change of F1 macro score with different k and smoothing parameter. The preferred combination is [k = 1, smoothing parameter = 0.5] and corresponding Hamming loss is 0.119, exact-match accuracy is 0.398. Comparing to results predicted with binary relevance algorithm, Hamming loss is larger and exact-match accuracy is lower, which might be limited by the complexity of the kNN

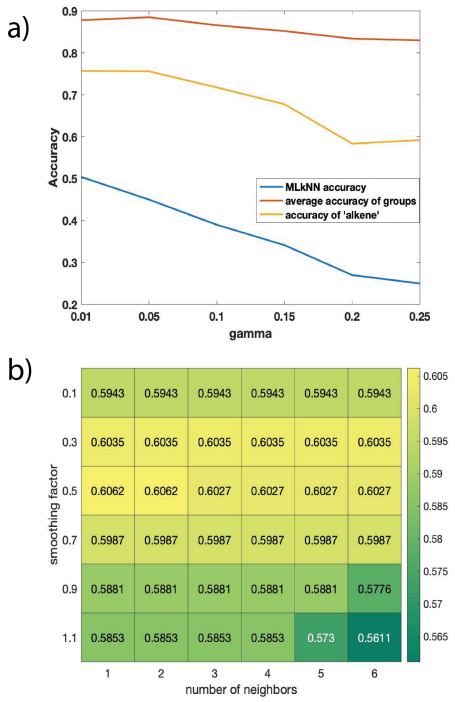

Figure 5. (a) Accuracy change with different gamma value [Red: average accuracy of 9 functional groups; Yellow: accuracy of 'alkene'; Blue: exact-match accuracy of binary relevance multi-classification model] (b) Grid search f1 score result with 5 fold cross validation on MLkNN algorithm

algorithm. Another possible reason is that there exists little relevance among different functional groups.

3.3. Peak shifting regression

Multi-labeled classification makes it possible for qualitative analysis of all kinds of organic compounds. However, we still can not extract any quantitative information from IR spectrum. If we could figure it out what

Functional group	1Cyclanes	2Esters	3Ether	4Alkene	5Alcohol	6Ketone	7Benzene	8Alkyne	9Acid
Training score	1.00	1.00	1.00	0.94	0.97	0.97	0.86	0.98	1.00
Testing score	0.76	0.91	0.93	0.83	0.88	0.93	0.84	0.98	0.95
Converged iteration	143	86	75	49	43	63	98	98	94

Table 4. Summarized results of multilayer perceptron (MLP) classification.

happened for certain type of molecules if the molecular length tend to be longer, we could then predict the length of molecules once we read the IR peak position. Here, to study the relation between peak position and molecular length, we computed DFT simulation results for Alkane from number of Carbon = 2 to 50. Since the peak position shift is really tiny, we can't accurately record the peak shift from IR spectrum in NIST, that's why we chose simulation results as research objects. We first simulated 49 Alkanes molecular structure with length of 2 carbons to 50 carbons. All the IR spectrum were computed and collected. 8 most distinguishable peaks were extracted form spectrum. 15 different regression methods were carefully conducted. Results of 6 typical regression methods displayed in Table 5. Exponential Gaussian Process Regression does perfect work with root mean square error (RMSE) as 0.4474 and R-squared value as 1.00. It makes sense due to the thermodynamic property of electron. A IR signal could be generated if molecule absorb certain energy and cause structural vibration. In other words, we have dipole moment (molecular vibration lead to dipole movement) at some certain energy level. IR spectrum typically shows the probability of electrons who lied in different energy level. We could observe a strong peak if electron is in favor of some certain energy level with specific wavenumber. And for longer molecules will have a periodic increased number of electron The probability distribution follow "Boltzman Distribution", which is a form of exponential Gaussian distribution with the only variable n as number of electron. So once we use Exp-GPR to fit our data, it works perfectly. Actually, all the Gaussian related methods perform well. Linear regression results are also shown in the last column of the chart. Linear fitting is not efficient to plot the relation between peak shift and molecular growth. We also plotted the results and predicted data in Figure 6. Figure 6a exhibits perfect matching prediction with small residuals for exp-GPR model.Figure 6b and 6c depicted the results for Coarse SVM and linear regression. It is found that the error for the first several dots are really huge. Again, since the electron behavior obeys Boltzman distribution, the peak shift tend to be pretty larger for the first several molecules with few electron. That's why we can't use linear regression to get accurate relation between peak shift and molecular size growth. In summary, we managed to find a proper regression method to understand the relation between peak shift and molecule growth. It provide us a new angle to analyze IR spectrum as "semi-quantitative" characterization. We could not only extract useful peak information to classify which type of molecule it should be, but also we could predict molecular size by computing peak position shifting.

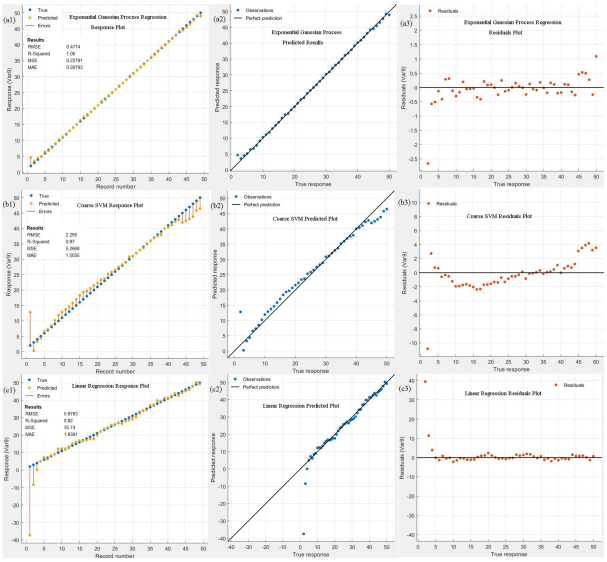

Figure 6. The regression results and related response plot (1), predicted plot (2) and residuals plot (3) for (a) Exponential Gaussian Process Regression; (b) Coarse SVM and (c) Linear Regression method.

4. Conclusions

Decision tree, SVM and MLP algorithms were implemented and compared for binary classification. Similar results were obtained from three algorithms. Functional group 1, 4 and 5 (cyclanes, alkene and alchohol) had the lowest accuracy with all three algorithms because many of their peak ranges overlap with each other, which makes the classification tasks more complicated. Linear SVM is found to be a suitable method to initialize classification for carbon-rich structure while rbf SVM seems to work better to study the classification problems for oxygen-rich samples. Two different multi-label classification methods were used in the paper: binary relevance with SVM as base classifier and MLkNN. The

	Exp-GPR	RQ-GPR	Coarse SVM	Boosted Trees	Stepwise LR	LR
RMSE	0.4774	1.3179	2.2952	2.7415	2.8976	5.9784
R-squared	1.00	0.99	0.97	0.96	0.96	0.82

Table 5. Results and performance indicator of multiple regression method. Exp-GPR:Exponential Gaussian Process Regression; RQ-GPR:Rational Quadratic Gaussian Process Regression; LR:Linear Regression; RMSE:Root Mean Square Error

optimized parameters were used in the binary relevance classifier. The results show that binary relevance approach does better job on multi-label spectrum classification task with higher exact-match accuracy (49%) and lower Hamming loss (10%). Finally, a molecular length-based regression were conducted. the relation between IR peak shift and molecular length growth were carefully studied. A perfect fitting model was obtained with exponential Gaussian process regression methods. We could rational compute and predict the length of Alkanes molecules once we found out the peak shift for certain IR signal. It provides us a new angle to analyze IR spectrum as "semi-quantitative" characterization. It will further achieve the development of IR analysis to be total molecular analysis with qualitative and quantitative information.

References

[1] F. De Comité, R. Gilleron, and M. Tommasi. Learning multi-label alternating decision trees from texts and data. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 35–49. Springer, 2003.

[2] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *Advances in neural information processing systems*, pages 681–687, 2002.

[3] D. I. Ellis, D. Broadhurst, S. J. Clarke, and R. Goodacre. Rapid identification of closely related muscle foods by vibrational spectroscopy and machine learning. *Analyst*, 130(12):1648–1654, 2005.

[4] T. Howley, M. G. Madden, M.-L. O'Connell, and A. G. Ryder. The effect of principal component analysis on machine learning accuracy with high dimensional spectral data. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 209–222. Springer, 2005.

[5] Y. Langeron, M. Doussot, D. J. Hewson, and J. Duchêne. Classifying nir spectra of textile products with kernel methods. *Engineering Applications of Artificial Intelligence*, 20(3):415–427, 2007.

[6] P. Lasch, M. Diem, W. Hänsch, and D. Naumann. Artificial neural networks as supervised techniques for ft-ir microspectroscopic imaging. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 20(5):209–220, 2006.

[7] R. Messenger and L. Mandell. A modal search technique for predictive nominal scale multivariate analysis. *Journal of the American statistical association*, 67(340):768–772, 1972.

[8] J. R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[9] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 254–269. Springer, 2009.

[10] M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In *IEEE international conference on neural networks*, pages 586–591. IEEE, 1993.

[11] R. E. Schapire and Y. Singer. Boostexter: A boosting-based system for text categorization. *Machine learning*, 39(2-3):135–168, 2000.

[12] S. Stein. Infrared spectra in nist chemistry webbook, nist standard reference database number 69, ed. pj linstrom and wg mallard, national institute of standards and technology, gaithersburg md, 20899.

[13] A. Subasi. Classification of emg signals using pso optimized svm for diagnosis of neuromuscular disorders. *Computers in biology and medicine*, 43(5):576–586, 2013.

[14] P. Szymański and T. Kajdanowicz. A scikit-based Python environment for performing multi-label classification. *ArXiv e-prints*, Feb. 2017.

[15] T. Van Gestel, J. A. Suykens, B. Baesens, S. Viaene, J. Vanthienen, G. Dedene, B. De Moor, and J. Vandewalle. Benchmarking least squares support vector machine classifiers. *Machine learning*, 54(1):5–32, 2004.

[16] V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[17] M.-L. Zhang and Z.-H. Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.

[18] J. Zhao, Q. Chen, X. Huang, and C. Fang. Qualitative identification of tea categories by near infrared spectroscopy and support vector machine. *Journal of Pharmaceutical and Biomedical Analysis*, 41(4):1198–1204, 2006.

[19] T. Zou, Y. Dou, H. Mi, J. Zou, and Y. Ren. Support vector regression for determination of component of compound oxytetracycline powder on near-infrared spectroscopy. *Analytical Biochemistry*, 355(1):1, 2006.