AJCAI 2022 Paper 3476: Response Letter

Jesse Wood[1][0000−0003−3756−2122], Bach Hoai Nguyen[1][0000−0002−6930−6863], Bing Xue[1][0000−0002−4865−8026], Mengjie Zhang[1][0000−0003−4463−9538], and Daniel Killeen[2][0000−0002−4898−6724]

[1] Victoria University of Wellington, Te Herenge Waka, PO Box 600, Wellington 6140, New Zealand
{jesse.wood, bach.nguyen, bing.xue, mengjie.zhang}@ecs.vuw.ac.nz
[2] Plant and Food Research, Port Nelson, Nelson 7010, New Zealand
daniel.killeen@plantandfood.co.nz

1 Review I

SCORE: SCORE: 1 (weak accept)

The first reviewer had these general comments for the paper.

> This paper automates processing of raw Gas Chromatography data to classify biomass that include fish species and fish body parts. First, this paper proposes a preprocessing imputation method to align timestamps in the training data. Then, it uses various machine learning methods to develop models for the classification tasks. Experimental results show that the SVM approach performs the best, however, visitation shows not all the features are needed. So this paper also uses four existing feature selection methods.
> This paper is well-written and easy to read. It explains the motivations behind the work and explain why a machine learning method is needed. The reason is the manual approach needs human experts and is expensive and time-consuming. This paper is mainly about using existing machine learning algorithms in an application. The experimental setup for both classification and then feature selection is good. The experimental results are good as well.

1.1 Literature Review

Comment:

> What I found missing is a literature review on the problem or similar problem. If this is the first of such work, the authors could explicitly claim that. Otherwise, discuss existing methods and perhaps compare with those as well.

Fixes:

1. Introduction adds references to similar existing methods for classification of GC data in food science, and the limitations of this work. This motivates the interpretable Linear SVM proposed in this work.
   [Introduction]
   Previous works using CNNs, *[1,3]*, showed high classification accuracy on gas-chromatograph data. However, these black-box models do not produce interpretable models, making it difficult to verify/troubleshoot for fish processing in a factory setting.
2. Section 3 on preprocessing explicitly adds references to similar existing methods.
   [3 Preprocessing]
   The authors are aware that there are more complex methods for imputing missing values, *[5,6]*, but they are not the focus of the paper and will be left for future work [...].

1.2 Imputation Contribution

Comment:

Although data imputation is an important part in the pipeline, however just 0 filling, while that makes sense, is not really a contribution. The time alignment in the data appears to be trivial as well,

Fixes:

1. The abstract removes reference to the filling 0 imputation.
   [Abstract]
   ~~Firstly, the paper proposes a preprocessing imputation method for aligning timestamps in Gas Chromatography data.~~
2. The introduction lightens the impact of the imputation contribution,
   [Introduction]
   The paper ~~proposes~~ *finds* an effective method to detect and fill the missing packets/features which ~~significantly~~ improves the classification performance over using the raw data.
3. Section 3 on preprocessing explicitly adds references to more advanced techniques to highlight the trivial nature of filling 0 for aligning GC data.
   [3 Preprocessing]
   The authors are aware that there are more complex methods for imputing missing values, *[5,6]*, but they are not the focus of the paper and will be left for future work [...].

1.3 Figure Formatting

Comment:

This figures need to be larger and visible.

Fixes:

1. Figure 1, the gas chromatograph, was enlarged to 0.8 of the linewidth.
2. Figure 2, the hyperplane coefficients, were changed to be two figures stacked vertically, both 0.8 of the linewidth.
3. Figures 3,4, the feature selection results, cannot be made any larger without exceeding the page limit (unless there is some black magic I am unaware of).

2 Reivew II

SCORE: -1 (weak reject)

The second reviewer gave the paper these general comments.

> This paper provides an interesting application of ML for fish classification using fatty acid Chromatographic data. It proposes a pre-processing imputation method for aligning timestamps in Gas Chromatography data, it demonstrates SVM could classify compositionally diverse marine biomass based on raw chromatographic fatty acid data, which can highlight important features for classification, and it also demonstrates that feature selection reduces dimensionality and improves classification performance by accelerating the classification system by four times.

2.1 Preprocessing Experimental Results

Comment:

> However, the motivation and research problem is not clear; for example, you need to demonstrate your pre-processing works using experimental results.

Fixes:

1. The model intends to be deployed in a factory setting for fish processing. Therefore an interpretable and accurate model is required. The paper had a section added to the introduction to clarify the constraints of this application. This concretizes the real-world applicability and scope of the research problem and motivates the need for an interpretable model.
   > [Introduction]
   > However, fatty acid data must be carefully processed and interpreted by domain experts (i.e. chemists), which is very expensive and time-consuming. *Previous works using CNNs, [1,3], showed high classification accuracy on gas-chromatograph data. However, these black-box models do not produce interpretable models, making it difficult to verify/troubleshoot for fish processing in a factory setting.*
2. **TODO:** KNN classification results for imputation methods.

2.2 Contributions

Comment:

Also, no innovative techniques have been developed, so the contribution is not enough; you should provide a new method to compare with the methods in Tables 3 and 4.

Fixes:

1. Future work will likely extend this conference paper into a journal paper, where the authors will propose new methods for imputation, classification and feature selection. The work would compare techniques from evolutionary computation to these existing results as suggested by the reviewer.
2. The paper does not propose a new method, but rather uses existing methods to solve a new problem. The novelty of the paper is its application. The model intends to be deployed in a factory setting for fish processing. Therefore an interpretable and accurate model is required. The paper had a passage added to the introduction to clarify the constraints of this application:
   [Introduction]
   However, fatty acid data must be carefully processed and interpreted by domain experts (i.e. chemists), which is very expensive and time-consuming. *Previous works using CNNs, [1,3], showed high classification accuracy on gas-chromatograph data. However, these black-box models do not produce interpretable models, making it difficult to verify/troubleshoot for fish processing in a factory setting.*

3 Review III

SCORE: SCORE: 0 (borderline paper)

Comment:

Page 6: "The hyperplane is represented by a weight vector in which each weight is associated with a feature. The larger the weight, the more important the corresponding feature. After an SVM classification algorithm is trained on the training set, an SVM classifier containing a learned weight vector is obtained. This section analyses the learned weight vector to examine the contribution of each packet/feature."
SVMs implement kernel methods to transform original data items into a high dimensional feature space where the input samples become linearly or mostly linearly separable. SVMs can learn the hyperplane in the feature space, which separates the training data with the widest margin. The hyperplane is constructed in the feature space that is nonlinearly related to input space. The weight vector representing the hyperplane in the feature space wouldn't match the features of original data items

in input space neither in dimensions nor in physical significance. The hyperplane, when being mapped to input space, becomes irregular contours outlined by support vectors. The important features (with larger weight) in feature space can hardly have their corresponding features in input space.
How to use weight vector of the hyperplane in feature space to examine the contribution of each packet/feature in input space?

Response:

(Usually) a conventional SVM uses a non-linear kernel, and the sklearn library defaults to the radial basis function (RBF) [4]. However, the paper states that experiments use a linear SVM model,

1. [Introduction] Experiments find that kernel-based classifiers, particularly *linear SVM*, achieve high classification accuracy on the fish data [...]
2. [4.1 Experiment Settings] These experiments compare five well-known classifications: K Nearest Neighbours (KNN where K is set to 3), Naive Bayes (NB), Random Forest (RF), Decision Trees (DT), and *Linear Support Vector Machines* (SVM) [...]
3. [5.4 Feature Selection Methods] In this work, a *linear SVM* is used as the wrapped classification algorithm since it achieves good classification performance [...]
4. [5.3 Experimental Settings] For each method, the balanced classification accuracy is measured with a *linear SVM* classification algorithm [4] [...]
5. [Conclusion] Among the considered classification algorithms, *linear SVM* achieves the best classification performance since it is suited to high-dimensional problems [...]

A linear kernel performs a linear transformation, which preserves the distance between points, mapping each instance to a 4800-dimensional vector in the feature space, then creates a hyperplane to linearly separate the classes in that feature space. Therefore the hyperplane coefficients, given in section 4.3, correspond to the original features.

3.1 Time Complexity

Comment:

Page 9: Meanwhile, PSO can remove 75% features, which means the classification system can be four times faster given the number of required packets/features is reduced by four times.
Considering the dimension of features of input data, does the classification speed linearly vary with the reduction amount of features?

Response:

Again, the reviewer has assumed the SVM used a non-linear kernel, in which case the complexity for non-linear SVM is expected to be $O(|\mathbb{X}|^2)$, where $|\mathbb{X}|$ is the number of training instances [2]. On page 9 the paper claims a linear speed up inversely proportional to the feature reduction,

> Meanwhile, PSO can remove 75% features, which means the classification system can be four times faster given the number of required packets/features is reduced by four times.

The paper uses SVM with a linear kernel, with time complexity $O(|\mathbb{X}| \times f_n)$, where f_n is the number of features. Therefore a 75% feature reduction would speed up the classification by a factor of 4.

Fixes:

1. (*Potential?*) Include complexity analysis of the Linear SVM in the introduction to the feature selection.

   > [5.4 Feature Selection Performance on Fish Species Classification] *The linear SVM has time complexity $O(|\mathbb{X}| \times f_n)$, where f_n is the number of features.* Meanwhile, PSO can remove 75% features, which means the classification system can be four times faster given the number of required packets/features is reduced by four times.

References

1. Bi, K., Zhang, D., Qiu, T., Huang, Y.: Gc-ms fingerprints profiling using machine learning models for food flavor prediction. Processes **8**(1), 23 (2020)
2. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST) **2**(3), 1–27 (2011)
3. Matyushin, D.D., Buryak, A.K.: Gas chromatographic retention index prediction using multimodal machine learning. Ieee Access **8**, 223140–223155 (2020)
4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
5. Tomasi, G., Van Den Berg, F., Andersson, C.: Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. Journal of Chemometrics: A Journal of the Chemometrics Society **18**(5), 231–241 (2004)
6. Zhang, D., Huang, X., Regnier, F.E., Zhang, M.: Two-dimensional correlation optimized warping algorithm for aligning gc× gc- ms data. Analytical Chemistry **80**(8), 2664–2671 (2008)