Repsonse Letter

No Author Given

No Institute Given

Abstract. Keywords:

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

Their review raised there specific issues about the paper. The first regarding a missing literature review containing related work.

> 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.

1.2 Imputation Contribution

The second issue was including the imputation as a contribution for the a paper.

> 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,

1.3 Figure Formatting

The third was formating, the figures need to be larger and easier to read, the camera-ready instructions for the conference specify at least a dpi of 800.

> This figures need to be larger and visible.

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

The reviewer had two specific issues with the paper. The first was the motivation and reserach problem were not clear, with a particular focus on the lack of experimental results in our preprocessing section.

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

2.2 Contributions

The felt as if the paper contribution was not enough, our experiments should evaluate a new method and benchmark it against the existing results.

> 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.

3 Review III

SCORE: SCORE: 0 (borderline paper)

The third reviewer had two questions regarding my paper.

The first issue mentioned the SVM uses a non-linear kernal, so the hyperplane coefficients have no relation to input features. This would make the analysis and claim on an interpretable model in section 4.3 invalid.

> 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?

(Usually) a conventional SVM uses a non-linear kernel, and the sklearn library defaults to the radial basis function (RBF) [2]. However, the paper states a Linear SVM model is used,

1. Experiments find that kernel-based classifiers, particularly *linear SVM*, achieve high classification accuracy on the fish data [...]
2. 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. In this work, a *linear SVM* is used as the wrapped classification algorithm since it achieves good classification performance [...]
4. For each method, the balanced classification accuracy is measured with a *linear SVM* classification algorithm [2] [...]
5. 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.

3.1 Time Complexity

Their second asks if for the time complexity of the classifer would speed up linearly with the number of features.

> 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?

Again, the reviewer has assumed the SVM used a non-linear kernal, 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 [1]. 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 linear kernal, 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.

References

1. 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)
2. 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)