A comparative Study in Recognizing Patterns through Perceptrons and Support Vector Machines

Jamie Mariani, CFA, 2021 April

Abstract:

This study critically evaluated different pattern recognition machine learning algorithms performing supervised learning on symptoms of Chronic Obstructive Pulmonary Disease (COPD). The algorithms evaluated are Support Vector Machines (SVM) and Artificial Neural networks (ANN). Models were tested with a variety of different hyper parameters which were scored using a grid search. These models were then validated using stratified crossing. The best models from this investigation were then compared and evaluated by confusion matrices and Receiver Operation curves (ROC). For predicting classification of lung health (whether a patient was healthy or had COPD). To carry on from previous research the healthy control group also included those with asthma to allow to see the effect of noise on the prediction accuracy. The study found that the SVM model worked better than the ANN.

1. Introduction

Chronic Obstructive Pulmonary Disease (COPD) is an obstructive lung disease which causes long-term breathing problems (Roversi et al. 2017; Vogelmeier et al. 2017). It currently affects around 10% of the world's population and is the 4th major cause of mortality and morbidity (Mulhall and Criner 2016; Ho et al. 2019). Although in the developing world smoking has decreased significantly COPD cases are increasing annually (Mulhall and Criner 2016; Reitan and Callinan 2016) COPD causes a progressive and currently irreversible narrowing of airways, which causes shortness of breath (Mulhall and Criner 2016). The main causative factor is smoking. There is currently no cure for COPD but drugs and early diagnosis can help cease further damage, identify the right palliative care and improve life quality (Ho et al. 2019; Mulhall and Criner 2016).

The current test for COPD utilises spirometry which measures lung function specifically volume and airflow speed (Arne et al. 2010). Spirometry isn't incredibly accurate. One study confirmed only a third of Swedish patients diagnoses and other research showed accuracy of around 64.5–79.9% (Arne et al. 2010). Sputum (mucus from the lower airways) production is another symptom of COPD and may be the answer to accurate COPD diagnosis and monitoring (Vogelmeier et al. 2017). Saliva can be analysed as a biomarker to look at how the condition progresses. To do this permittivity biosensors are used on saliva to characterize it's dielectric properties (Zarrin et al. 2020).

There are multiple factors which determines susceptibility to COPD. From smoking habits, age and gender. Some research suggests women who smoke may be disproportionality effected (Chapman et al. 2001). Supervised machine learning algorithms could be a great option for analyse COPD data. In this study SVM and MLP algorithms will be compared and see if they can accurately predict if an individual has COPD. In section 2 the data set will be briefly explained. Section 3 will focus on the models and methods used. Section 4 will evaluate data and then section 5 will contain the conclusions of the study.

1.1. Support vector machines

SVMs are machine learning algorithms which can be used for classification or regression. Literature points to SVMs preforming well with biological data sets. in the past they have been used analysing gene expressions profiles, diagnosing a disease and looking at structures and functions of proteins (Ben-Hur et al. 2008). Their ability to perform well on biological data sets is due to their non-linear nature which can be applied using kernels. The kernels in SVMs give them the ability to theoretically work with any number of hyper dimensions (Schölkopf and Smola 2003). This allows them to identify decision boundaries which optimise the margin distance between different data classes. Meaning they can handle complex biological data sets which couldn't be spate accurately using purely linear models. This is one definite advantage for using them over ANN as its allows us to work with data which a purely linear model couldn't. SVMs are able to. The disadvantages of using SVM is they can struggle with large data sets due to long training times which can be more computationally intensive, this is not an issue in this study but should be considered when doing future research.

1.2 Artificial neural networks

ANN are a machine learning algorithm comprised of a collection of connected units called nodes. ANN have 3 layer types which handle data the input layer which takes the inputted data, hidden layers which then process the data and then finally the output layer which give the output of the model. ANN use their layers of nodes to try and replicate the biological neural networks found in the brains of animals (Chen et al. 2019).

What's great about ANN is that they are highly modifiable through backpropagation. The loss or gradient descent seen in a model can be reduced through modifying the learning parameters. This is great as it allows you to fit your data well but can lead to overfitting if the model becomes too specific to the training data (Lawrence and Giles 2000).

1.3 Hypothesis

The SVM and ANN binary model classifiers will be better at predicting COPD than spirometry.

2. Data

The data set used in this study comes from the UCI machine learning repository (Zarrin 2020). The data is from a clinical study looking into COPD diagnosis. It contains 399 samples of which 90 were used. Of the 90 samples 40 came from COPD patients, 40 from healthy controls and 10 from people suffering asthma. There were 8 attributes, which were a mix of both continuous and discrete data. The strings in the data set were turned into numerical values as follows. Diagnosis label was changed to 0 = COPD and 1 = Healthy control or asthma. Data from COPD patients with infections were discarded to follow GOLD guidelines for COPD diagnosis on top of this acute respiratory infection can change saliva's properties (Mirza et al. 2018). Smoking status was displayed on the data numerically as 1=Non-smoker, 2=Ex-smoker, 3=Active-smoke. Gender was displayed with 1=Male and 0=Female. There was imbalanced data for gender, 33 of the samples were from females while the remaining 57 come males. It is still to be confirmed if gender plays a significant role and due to the sample numbers not being hugely different techniques such as SMOTE or Nearmiss Algorithms weren't used to balance the data. The age attribute was already numerical so not changed. The continuous data included both real and imaginary parts of the permittivity of the saliva. The average and minimum values were used. The real part data shows the energy absorption known as its dielectric properties. The imaginary part measured the saliva's energy loss.

2.1 Initial data analysis

To begin the initial data analysis a heat map of the data was generated to visualise correlations within the data. This can be seen in Figure 1. The data was visualised to help identity any hidden patterns in the data which cannot be seen from just looking at the dataset. Heat maps help to identify what attributes in the dataset may influence the models. As there is a mix of both continuous and discrete data a heat map was a superior way of presenting the data. Instead of using a violin plot for the continuous data and then a pairwise wise comparison on a histogram for discrete data.

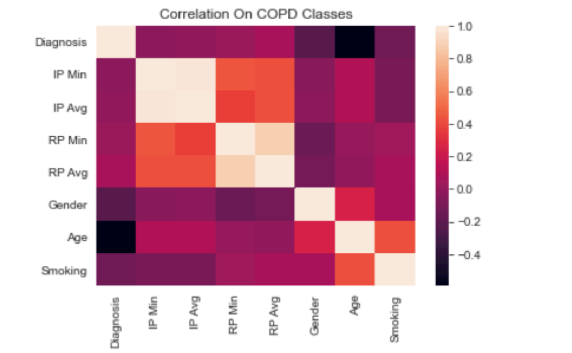

Figure 1. Heat map for attributes in the data set

From figure one we can see that the real part average shows a strong positive correlation with diagnosis. This points to dialect properties averages (real part average) being the main determinant factor for diagnosing COPD from biosensor analysed saliva samples. From the heat map, it can also be seen there is a strong negative correlation between age and a healthy diagnosis, showing older people are more likely to be diagnosed with COPD. There is also a weaker negative correlation between smoking and being classed as healthy. The data also points to their being a link between being male and getting COPD, this may be bias in the data set but also could be due to men using more tobacco product than women.

Figure 2 helps us to visualise the distribution of the data for the real part average. As we can see the data shows quite a lot of overlap. As real part average has been identified as a determinant function in the heat map it is important to understand its distribution so we can pick models which best suit the data. This overlap shows that a linear classier may struggle to deal with the data. Figure 2 also helps illustrate how infection can skew the data.

3. Methods

Now the details of how the SVM and ANN models were trained and validated. The architecture and hyper parameters used to build the models will also be explained.

3.1. Methodology

20% of the data was withheld from the original data set as the testing set. The remaining 80% of the data was used to train the model. To ensure that the comparisons between the SVM and ANN models were fair a random state parameter set to 21 was used to ensure the shuffling of the dataset for different models was kept constant.

Model selection was done using a grid search to optimise the hyper parameters of the SVM model. For training and validation of the data the models were ran through a 5-fold stratified cross-validation. The 5-fold stratified cross-validation to estimate the skill of the models on new data. It helps to remove bias of the model's skills which can't be done with a train test split. Due to the dataset size, it would have been unwise to a K fold higher than 5. When the k mean is too large only a small number of sample combinations are possible. This would limit the number of iterations that are unique, thus running the risk of duplicates

3.2 Architecture and Parameters used for the SVM

SVMs are a non-perceptron based machine learning classifier. They also allow you to only select a subset of the data the SVM for classification. Due to SVMs not using perceptrons they don't have the issue of getting stuck in random local minima. Although they do not start with random weights they still must be trained. The kernel, soft margin constant and the gamma were all adjusted to help build a better model. Due to the non-linear nature of the data seen in Figure 2 a radial basis function (RBF) kernel was used. This is as RBF allows you to transform the data and view it in higher dimensions in a linear manner, then report it back to the original data as a non-linear separator. As an RBF is used this introduces the Gamma hyper parameter which controls the smoothness of the classification boundaries or the inverse of radius of influence of the SVMs used for classification. High values reduce smoothness and decrease decision function margins which can lead to more accurate models, but also overfitting. While lower values lead to smoother models which increases the margin. C or the soft margin constant works as regularization parameter in SVMs to control error. Setting this parameter comes with a trade-off between correct classification of training data or maximisation of the decision functions margin. The larger the C the smaller the margin will be thus higher accuracy, while a smaller C will have a lower accuracy and a larger margin.

3.3 Architecture and Parameters used for the ANN

For the ANN a logistic regression a Sigmoid output function was used. This is as the Sigmoid function is great for classification problems as it transforms the input value to either 0 or 1. A standard scaler was used to scale the features of the model so that unit variance and mean of zero. BCELoss criterion was used. It measures the Binary Cross Entropy between the target and the output. Stochastic gradient descent was used for optimization. This means it helped reduce as the loss of a predictive model with respect to the training data.

The testing stage involved 5 validation checks and a maximum number of epochs of 40. The epochs were limited as although loss went down for each epoch validation the model started to over fit the training data.

4.1 Model selection

The best SVM model was selected using a grid search it had an accuracy of 94%. The model had a C of 10 and a gamma of 0.0001. The best ANN accuracy was 55%. To test fir the best SVM's ability to generalise the predictions of the model a cross validation process was used. This involved finding the loss by comparing the predicted labels generated by the model and the actual labels. In the training process for the SVM the gridsearch showed that a C of 10 and a Gamma of 0.001 were best for prediction. The high C indicates that the model has low decision function margins but high accuracy which can be seen from the prediction accuracy of 94%. Before the training process the accuracies were only 55% for the test and training data, by fine tuning the hyper parameters a higher accuracy was achieved. At the end of the process the average 5-fold cross validation was 89% for the test data. A 5-fold cross validation showed an average accuracy of 85%.

4.2 Algorithm comparison

For use in diagnosis the SVM model was superior. This was most likely due to it being more complex than the ANN.

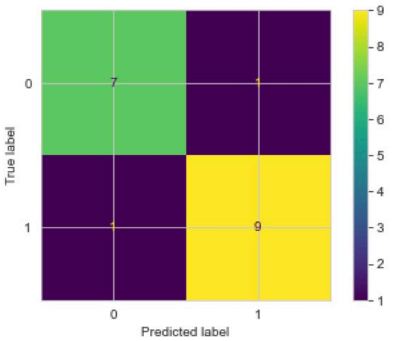

Figure 2. Confusion Matrix for test data ran through the SVM model

Figure 2 shows that the SVM model has a very low false negative which is important when diagnosing someone with COPD. As COPD cannot be cured and only treated. Spotting it early is vital for ensuring a higher quality of life to COPD sufferers.

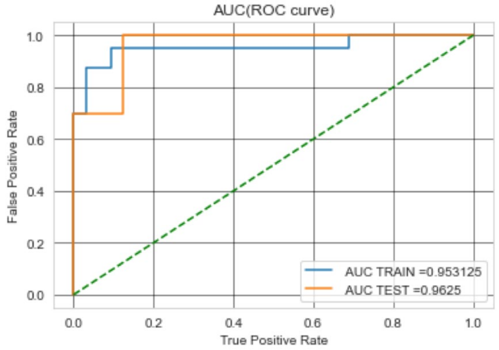

Figure 3. Receiver Operation Curve for the training and test data sets using the SVM model

ROC is a great measure to look at the accuracy between training and test sets. This is as AUC which is the space under the curve is not effected by the number of samples. This allows for unbalanced datasets to be compared without bias.

5. Conclusion

This study investigated the accuracy of a SVM and MLP of diagnosing COPD from a mix of discrete and continuous data. The conclusion is that only the SVM model could be a viable option for diagnosing COPD from biosensor analysed saliva data and discrete patient background data. The SVM was able to predict accurately even with Asthma suffers included in the healthy group, showing the model can deal with noise. The maximum accuracy if the SVM model was 94% this is compared to the 55% achieved by the ANN. From this we can see that the SVM model outperforms spirometry, the current diagnosis. The maximum reported accuracy of spirometry is 79.9% but as before mentioned in some clinical settings only a third of patients can be diagnosed, so there is a high variability in accuracy. The ANN model was not as good but this is most likely due to the simplicity of the model. In the end the hypothesis was proven to be partially true.

This study showed the ability power of ROC curves for evaluating classifiers, measuring well their accuracies independently from the proportions of classes in the dataset and also giving a visual idea of the sensitivity and sensibility of the classifier model.

Future investigations should look at different model’s accuracy of diagnosing COPD. One interesting model to try could be XGBoost algorithm. This non perceptron machine learning algorithm would be interesting to investigate as it is designed to handle both discrete and continuous data (Chen et al. 2017). Using a XGBosst algorithm could lead to a model with a higher accuracy. Future research would also benefit from larger datasets, SVMS are great with small data sets but it would be interested to see how accuracy and computational load would change with a larger dataset, albeit that may need for progress in the ease of analysing saliva data.

References:

Arne, M. et al. (2010). How often is diagnosis of COPD confirmed with spirometry? Respiratory Medicine, 104(4), pp.550–556.

Ben-Hur, A. et al. (2008). Support Vector Machines and Kernels for Computational Biology F. Lewitter, ed. PLoS Computational Biology, 4(10), p.e1000173.

Chapman, K.R., Tashkin, D.P. and Pye, D.J. (2001). Gender Bias in the Diagnosis of COPD. Chest, 119(6), pp.1691–1695.

Chen, W. et al. (2017). Radar emitter classification for large data set based on weighted-xgboost. IET Radar, Sonar & Navigation, 11(8), pp.1203–1207.

Chen, Y.-Y. et al. (2019). Design and Implementation of Cloud Analytics-Assisted Smart Power Meters Considering Advanced Artificial Intelligence as Edge Analytics in Demand-Side Management for Smart Homes. Sensors (Basel, Switzerland), 19(9).

Ho, T. et al. (2019). Under- and over-diagnosis of COPD: a global perspective. Breathe, 15(1), pp.24–35.

Lawrence, S. and Giles, C.L. (2000). Overfitting and neural networks: conjugate gradient and backpropagation. In Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium. Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium. Como, Italy: IEEE, pp. 114–119 vol.1. [online]. Available from: http://ieeexplore.ieee.org/document/857823/ [Accessed April 7, 2021].

Mirza, S. et al. (2018). COPD Guidelines: A Review of the 2018 GOLD Report. Mayo Clinic Proceedings, 93(10), pp.1488–1502.

Mulhall, P. and Criner, G. (2016). Non-pharmacological treatments for COPD: Non-pharm COPD. Respirology, 21(5), pp.791–809.

Reitan, T. and Callinan, S. (2016). Changes in Smoking Rates Among Pregnant Women and the General Female Population in Australia, Finland, Norway, and Sweden. Nicotine & Tobacco Research, p.ntw188.

Roversi, S., Corbetta, L. and Clini, E. (2017). GOLD 2017 recommendations for COPD patients: toward a more personalized approach. COPD Research and Practice, 3(1), p.5.

Schölkopf, B. and Smola, A.J. (2003). A Short Introduction to Learning with Kernels. In S. Mendelson & A. J. Smola, eds. Advanced Lectures on Machine Learning. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 41–64. [online]. Available from: http://link.springer.com/10.1007/3-540-36434-X_2 [Accessed April 7, 2021].

Vogelmeier, C.F. et al. (2017). Global Strategy for the Diagnosis, Management and Prevention of Chronic Obstructive Lung Disease 2017 Report: GOLD Executive Summary. Respirology, 22(3), pp.575–601.

Zarrin, P.S., Roeckendorf, N. and Wenger, C. (2020). In-Vitro Classification of Saliva Samples of COPD Patients and Healthy Controls Using Machine Learning Tools. IEEE Access, 8, pp.168053–168060.

Appendix 1: Glossary

Term	Definition
Artificial Neural networks	Computing systems vaguely inspired by the biological neural networks that constitute animal brains
Asthma	Asthma is a condition in which your airways narrow and swell and may produce extra mucus
Binary	is a base-2 number system invented by Gottfried Leibniz that's made up of only two numbers: 0 and 1.
Binary Cross Entropy	it is a Sigmoid activation plus a Cross-Entropy loss
Biomarker	measurable indicator of some biological state or condition
Biosensors	an analytical device, used for the detection of a chemical substance, that combines a biological component with a physicochemical detector
Confusion matrix	able that is often used to describe the performance of a classification model on a set of test data for which the true values are known
COPD	is a chronic inflammatory lung disease that causes obstructed airflow from the lungs
Dielectric properties	molecular property that is fundamental in all the materials that are capable of impending electron movement resulting in polarization within the material on exposure to an external electric field
Gamma	defines how far the influence of a single training example reaches
Generalise	more widespread or widely applicable
Perceptron	an algorithm used for supervised learning of binary classifiers
Permittivity	a measure of the electric polarizability of a dielectric.
ROC	a graph showing the performance of a classification model at all classification thresholds
Sigmoid	Activation function which transforms values into a value between 0 or 1
Soft margin	A SVM margin which the distance from the decision surface to the closest data point
Spirometry	common test used to assess how well your lungs work by measuring how much air you inhale, how much you exhale and how quickly you exhale
Sputum	a mixture of saliva and mucus coughed up from the respiratory tract
Stochastic gradient descent	an iterative method for optimizing an objective function with suitable smoothness properties
Stratified crossing validation	The splitting of data into folds may be governed by criteria such as ensuring that each fold has the same proportion of observations with a given categorical value, such as the class outcome value.
Support Vector Machines	supervised learning models with associated learning algorithms that analyze data for classification and regression analysis
XGBoost	open-source software library which provides a gradient boosting framework

Appendix 2: Code implementation

SVM model building

The relevant packages were loaded to build the model. For building this particular SVM Scikit learn was the most important.

1. Load in data from ExasensA.csv

2. Clean data
-First the unneeded columns were removed, in this case only 'ID' was removed as it would have to significant effect on the model.
-Then rows with no data from the biosensor were removed as we wanted to look at its ability to predict COPD
-Labels for COPD and Asthma were turned to 1 and labels corresponding to COPD were turned to one so they could be read by the model and so that it could work as a binary classifier.

3. early data analysis
-A heat map was generated using Seaborn to look at correlations not visible from viewing the data set alone.
-A box plot was also generated to show the distribution of the Real part imagery data as this data showed the strongest correlation out of all the biosensor data classes.

4. Preparing test and training data
The training and test data is split up using test train split from sklearn

5. SVM training
-An untrained SVC() is ran using the training data and scored to get initial idea of how well a SVM works on the data. This model scored 55%
-The number of support vectors were then identified using the len() on svm.support_)
-The default rbf kernel was used
-GridsearchCV was used to search for the best C and gamna parameters
-gamma ranged from 0.00001, 0.001, 0.01, 0.1 while c ranged from 0.1, 1, 10
-The best parameters were found by using best_params_ on the gridsearch variable (searcher in this case). These were 0.0001 for gamma and 10 for C.
-The new SVC() with the best parameters was scored and gave an accuracy of 94%

6. Data validation
-The 5 K fold was used over a 10 k fold due to the small size of the data set.

ANN

1. Arne, M. et al. (2010). How often is diagnosis of COPD confirmed with spirometry? *Respiratory Medicine*, 104(4), pp.550–556.

Ben-Hur, A. et al. (2008). Support Vector Machines and Kernels for Computational Biology F. Lewitter, ed. *PLoS Computational Biology*, 4(10), p.e1000173.

Chapman, K.R., Tashkin, D.P. and Pye, D.J. (2001). Gender Bias in the Diagnosis of COPD. *Chest*, 119(6), pp.1691–1695.

Chen, W. et al. (2017). Radar emitter classification for large data set based on weighted-xgboost. *IET Radar, Sonar & Navigation*, 11(8), pp.1203–1207.

Chen, Y.-Y. et al. (2019). Design and Implementation of Cloud Analytics-Assisted Smart Power Meters Considering Advanced Artificial Intelligence as Edge Analytics in Demand-Side Management for Smart Homes. *Sensors (Basel, Switzerland)*, 19(9).

Ho, T. et al. (2019). Under- and over-diagnosis of COPD: a global perspective. *Breathe*, 15(1), pp.24–35.

Lawrence, S. and Giles, C.L. (2000). Overfitting and neural networks: conjugate gradient and backpropagation. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*. Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium. Como, Italy: IEEE, pp. 114–119 vol.1. [online]. Available from: http://ieeexplore.ieee.org/document/857823/ [Accessed April 7, 2021].

Mirza, S. et al. (2018). COPD Guidelines: A Review of the 2018 GOLD Report. *Mayo Clinic Proceedings*, 93(10), pp.1488–1502.

Mulhall, P. and Criner, G. (2016). Non-pharmacological treatments for COPD: Non-pharm COPD. *Respirology*, 21(5), pp.791–809.

Reitan, T. and Callinan, S. (2016). Changes in Smoking Rates Among Pregnant Women and the General Female Population in Australia, Finland, Norway, and Sweden. *Nicotine & Tobacco Research*, p.ntw188.

Roversi, S., Corbetta, L. and Clini, E. (2017). GOLD 2017 recommendations for COPD patients: toward a more personalized approach. *COPD Research and Practice*, 3(1), p.5.

Schölkopf, B. and Smola, A.J. (2003). A Short Introduction to Learning with Kernels. In S. Mendelson & A. J. Smola, eds. *Advanced Lectures on Machine Learning*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 41–64. [online]. Available from: http://link.springer.com/10.1007/3-540-36434-X_2 [Accessed April 7, 2021].

Vogelmeier, C.F. et al. (2017). Global Strategy for the Diagnosis, Management and Prevention of Chronic Obstructive Lung Disease 2017 Report: GOLD Executive Summary. *Respirology*, 22(3), pp.575–601.

Zarrin, P. (2020). Exasens Data Set. *UCI Machine learning repository*. [online]. Available from: https://archive.ics.uci.edu/ml/datasets/Exasens.

Zarrin, P.S., Roeckendorf, N. and Wenger, C. (2020). In-Vitro Classification of Saliva Samples of COPD Patients and Healthy Controls Using Machine Learning Tools. *IEEE Access*, 8, pp.168053–168060.