iCBST 2014 International Conference on Business, Science and Technology which will be held at Hatyai, Thailand on the 25th and 26th of April 2014.

Australian Journal of Basic and Applied Sciences, 8(5) Special 2014, Pages: 285-293

AENSI Journals

Australian Journal of Basic and Applied Sciences

ISSN:1991-8178

Journal home page: www.ajbasweb.com

Comparing Classification Techniques using Different Methods of Data Splitting for Artemisinin Compounds

[1]Rosmahaida Jamaludin, [1]Mohamed Noor Hasan, [2]Mohd Zuli Jaafar

[1]*Department of Chemistry, Faculty of Science Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia.*
[2]*Faculty of Applied Science, Universiti Teknologi MARA, Kampus Kuala Pilah,72000 Kuala Pilah, Negeri Sembilan, Malaysia.*

ARTICLE INFO

Article history:
Received 25 January 2014
Received in revised form 12 March 2014
Accepted 14 April 2014
Available online 25 April 2014

Key words:
Artemisinin, Chemometrics, Classification, Kennard-Stone, Duplex, Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Linear Vector Quantization (LVQ), Quadratic Discriminant Analysis (QDA).

ABSTRACT

In this study, Artemisinin compounds that were classified according to their anti-malarial activity values were used as a data set to develop predictive classification models namely Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Linear Vector Quantization (LVQ) and Quadratic Discriminant Analysis (QDA). The influence of Duplex and Kennard-Stone method of splitting data into a training and test set were also investigated. Their performance were evaluated based on the percent correctly classified (%CC) of both training and test set. Standardization was used as data pre-processing technique. The generated classification models have shown that Kennard-Stone data splitting technique produced higher percent correctly classified of test set in all models. Meanwhile, LDA approach was found to be superior with lower risk of over-fitting for artemisinin data set.

To Cite This Article: Rosmahaida Jamaludin, Mohamed Noor Hasan, Mohd Zuli Jaafar., Comparing Classification Techniques using Different Methods of Data Splitting for Artemisinin Compounds. ***Aust. J. Basic & Appl. Sci.,*** *8(5): 285-293, 2014*

INTRODUCTION

Malaria is an infectious and deadly disease and is a major health problem especially in the developing world. Artemisinin or *qinghaosu* as displayed inFigure 1, extracted from *Artemisia annua L.* has a significant anti-malarial activity. This powerful anti-malarial drug and its derivatives such as artemether, arteether, artesunate and dihydroartemisinin are effective measures against drug-resistant malarial parasites and rapid progression of malarial illness (Ploypradith, 2004). This sesquiterpene endoperoxide lactone has unique structure and mechanism of action characterized by the presence of a peroxide bridge, a known source of oxygen-free radicals that is responsible for its anti-malarial activity (Meshnick, 2002).Many ongoing studies have provided a better understanding of its molecular mode of action and the probable intermediates crucial for the anti-malarial activity (Cazelles, Robert, & Meunier, 2001; Kamchonwongpaisan & Meshnick, 1996; Olliaro, Haynes, Meunier, & Yuthavong, 2001; Posner, Cumming, Ploypradith, & Oh, 1995; Posner, *et al*., 1996). Continuing efforts have been made extensively to develop more potent, selective, nontoxic, clinically useful newer semi-synthetic and synthetic derivatives that are superior in activity in order to control this infection throughout the world**.**

Splitting the data into training and test set is widely used as validation method in pattern recognition in addition to model optimisation techniques such as cross validation and bootstrap. Different techniques of data splitting as well as data pre-processing and type of classification models used essentially influences the performance of the classification. Hence, in this work, we will investigate the performance of data splitting methods for classification to develop reliable predictive classification models and consequently compare them by utilizing well-characterized data set of artemisinin analogues collected by Avery (Avery, *et al*., 2002).

Corresponding Author: Rosmahaida Jamaludin, Department of Chemistry, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia.
E-mail: rosma@ic.utm.my

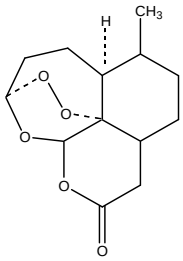


Fig. 1: Structure of artemisinin

The proposed strategy is to test the application of two different methods of data splitting to a series of compounds from the dataset using four classification techniques, namely Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Linear Vector Quantization (LVQ) and Quadratic Discriminant Analysis (QDA).Therefore, the aim of this study is to investigate the impact of different data splitting algorithms on the percent correctly classified of test and training sets and consequently, compare the performance of four classification techniques.

Experimentation:

The data set used in this study consisted of 197 compounds. Each of these compounds had its associated *in vitro* bioactivity values against the drug-resistant malaria strain *P. falciparum* (W-2 clone) expressed as log RA (relative activity)calculated from the experimentally derived control IC_{50} (reported in ng/ml)and corrected for molecular weight (*MW*) (Guha & Jurs, 2004).

$$\log RA = log\left(\frac{IC_{50}\, of Artemisinin}{IC_{50}\, of the analog}\right) \times log\left(\frac{MW of analog}{MW of Artemisinin}\right)$$

The classification are based on the anti-malarial response where compounds with log RA ≥ 0.00 were assumed as more potent analogues and those with log RA < 0.00 were considered as less potent analogues (Ferreira, *et al*., 2010). Therefore, for the purpose of modelling, a value of 1 and 2 were assigned to active and not active class respectively.

Initially, a total of 3764 descriptors which include topological, connectivity indices, constitutional indices, molecular properties, 2D-autocorrelation and Burden eigenvalues have been generated from three-dimensional representations of the compounds in the dataset. After the curation process which involves eliminating descriptors with zero values and zero variance, the number of descriptors was then reduced to 2075. Hence, the ratio of variables to samples is high and need to be handled properly to obtain reasonable assessment of predictive ability.

All the software packages used in this study were run using Microsoft Window XP on a Pentium IV system. ChemDraw Ultra version 6.0 (Cambridge Soft) was used to draw 2D model molecular structure of the compounds. Next, Chem3D Ultra version 6.0 was utilized to convert the molecular structure to 3D structure. The molecular descriptors for all the compounds were solely calculated using DRAGON software package Talete srl, DRAGON Version 6.0 http://www.talete.mi.it/(Todeschini, Consonni, Mauri, & Pavan, 2010). Further analyses were performed in Matlab 7.5.0 (2007) (The Mathworks Inc., Natick, MA).

Unsupervised data splitting methods used in this work were Duplex and Kennard Stone. Both approaches are based on molecular descriptors alone. The division into training and test set is recommended for the purpose of model building, model optimisation and independent validation to assess the quality of the predictive model (Richard G. Brereton, 2006). In other words, the whole data set of known class membership was divided into a training set to build the models and a test set of the remained compounds to test classification ability of the models or assess its predictive capability.The classification ability may be influenced by the choice of samples included in the training set, and thus, according to Capron (Capron X., Walczak B., Noord O.E. de, & D.L., 2005) these representative methods of selection of samples lead to better results. The algorithms used in this work were based on the related study of these two data splitting methods (Jaafar, 2011). The aim is to obtain more balanced and homogeneous training and test sets that evenly scattered over the whole space. The steps implemented for this technique include splitting the dataset into training and test set where normally 2/3 was selected as training set while the remaining in test set, equally distributed among each class, subjected to the limitations applicable to each procedure.As illustrated in Table 1, the training set contained approximately double the percentage of test set for both active and less active analogues while the number of molecules in test set of less potent analogues is slightly higher.

Subsequently, selected data pre-processingmethod namely standardizationwas applied on both training and then test setswhere all descriptors are centred on zero and have a standard deviation of one to give each variable equal importance. This approach is necessarysince the variables usedin this study have different units. The average value of a descriptor, $\bar{x}_j$is subtracted from each individual value, x_{ij}. Each descriptor value is then

divided by the standard deviation for that descriptor across all molecules, *I* as shown in the equation below. Consequently, each scaled descriptor then has variance of one.

$$^{std}x_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\frac{\sum_{i=1}^{I}(x_{ij}-\bar{x}_j)^2}{I}}}$$

Table 1: Number of compounds for both more and less potent analogues in training and test.

Artemisinin data set	More potent analogues (class 1)	Less potent analogues (class 2)
Artemisinin molecules	96	101
Training set	64	64
Test set	32	37

Both methods of data splitting used in this work only consider the values in the molecular descriptors. Their algorithms are listed sequentially.The following are the steps of Duplex algorithm (Jaafar, 2011).

Step 1: The Euclidean distances between all possible pairs of points are calculated.

Step 2: The points that are farthest away from each other are selected as training samples and removed from the data block.

Step 3: For the remaining samples, the points furthest apart from each other are put into the test set.

Step 4: When the samples assigned to each set are removed, the next iteration will be carried out until all the samples have been assigned to training and test set.

In this approach the sample is iteratively selected based on distance to ensure the maximum coverage of the data. The distance between two points, let say x_k and x_i is calculated using Euclidean distance as shown in the following equation (Puzyn, Mostrag-Szlichtyng, Gajewicz, Skrzyn, & Worth, 2011).

$$\boldsymbol{d_{x,kl} = \sqrt{(x_k - x_l)(x_k - x_l)'}}$$

The steps of Kennard-Stone algorithms (Jaafar, 2011) are listed below. As in Duplex algorithm, maximin criterion is used to find out which sample is the farthest one (Capron X., *et al*., 2005).

Step 1: Find the two point's $\boldsymbol{x_k}$and $\boldsymbol{x_l}$which are the most separated in the training set and removed these from the data.

Step 2: For the remaining data points, find the minimum distance of any two objects already selected in step 1.

Step 3: The training set is the set of samples which have the largest of these minimum distances.

Step 4: Iterates step 1 to 3 until the required numbers of training/test set samples are found.

The classification problems presented in this work focus on two class or binary classifiers also known as hard model in which samples are assigned to one of two groups. In other words, we attempt to classify artemisinin compounds into active and less active compounds. The process of classification is illustrated in Figure 2.

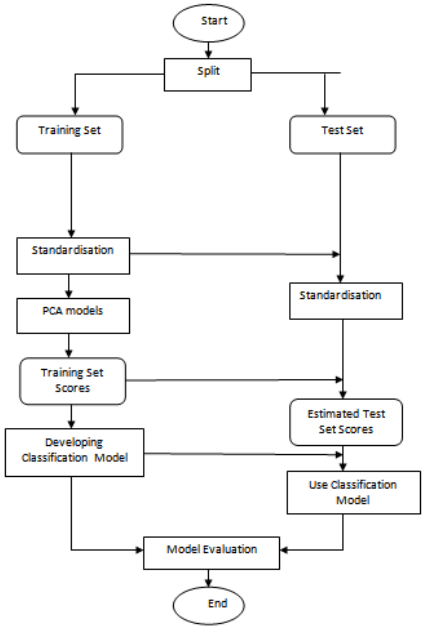


Fig. 2: The process of classification by using PCA scores

LDA is based on distance to each class centroid for two class problem such as Euclidean distance. However, the measured distance used in this study is Mahalanobis distance (*d*) as shown in the equation below which takes into account the variance or dispersion of each variables as well as correlation between variables by making use of the variance-covariance matrix (Richard G. Brereton, 2009).

$$\boldsymbol{d(i,g) = \sqrt{(x_i - \bar{x}_g)' C_P^{-1} (x_i - \bar{x}_g)}}$$

where x_i is the measurement obtained for the *i*th sample, $\bar{x}_g$is the centroid of class g and C_P^{-1} is the inverse of the pooled variance-covariance matrix acquired on the entire dataset and can be computed from the following equation:

$$\boldsymbol{C_p = \frac{(I_1 - 1)C_1 + (I_2 - 1)C_2}{(I_1 + I_2 - 2)}}$$

where I_g and C_g correspond to number of samples and variance–covariance matrix of class g respectively. Thus, the distance between a test sample of a given class centroid is weighted according to the overall variance of each variable. This type of classifier is suitable when the two classes have similar variance-covariance matrices, otherwise the pooled covariance matrix will be biased to the class with higher covariant matrix. The distance between a sample and the class centroid is computed and consequently each sample is assigned to the class with the smallest distance. Therefore, as the distance to a weighted centroid gets larger the probability of a sample belonging to a class decreases (De Maesschalck, Jouan-Rimbaud, & Massart, 2000).

QDA is another distance based classifier, an extended version of LDA classifier where the Mahalanobis distance to each class centroid is calculated using the sample variance-covariance matrix of each class individually where variance of each group is calculated separately instead of the overall pooled matrix as in the previous technique. Therefore, this type of boundary is more complex and no longer linear. The model is more suitable when the two classes have very different variance structures and will favour the class with high variance(Richard G. Brereton, 2009).Both LDA and QDA describe a simple decision boundary between two classes which has equal distance to the two centroids and the distance measures assume samples in a class are distributed as hyperellipsoids and variables in each class form a multi-normal distribution (Dixon & Brereton, 2009).

LVQ is based on a combination of straight lines to represent separation boundaries and uses codebook vectors to characterize each class in training set instead of original samples to form a model where the number of lines being a function of the number of codebooks. In this method, the decision boundary between the two classes is optimised by learning the position of the optimal codebook vectors that results in the smallest misclassification error (Dixon & Brereton, 2009). LVQ is a complex iterative procedure, partially sensitive to the samples selected for the training set, and the initial choice of codebooks. Since the codebooks are derived using an iterative neural network, the model will not be reproducible. The primary step for LVQ is to generate codebook vectors for each class in the training set and place them within their class region by using a sample that can be correctly classified. The distance between each sample to the nearest codebook is computed and consequently each sample is assigned to the class of the nearest codebook vectors.Increasing the number of initial codebook vectors per class can improve the description of the class boundary as well as reduce misclassification error but may cause overfitting and results in poor classification of test set (Richard G. Brereton, 2009). In this work, initial learning ratewas set to 0.09 and 10000 respectively(Lloyd, Brereton, Faria, & Duncan, 2007)while the number of codebook vectors were maintained at 3 to avoid biases of different classes' size by using the same number of codebook vectors in each class to construct the boundary. The class borders are approximately located after training of the codebook vectors is completed. Consequently, sample is assigned to the class of the nearest codebook.

Another useful and popular technique for classification is Support Vector Machine (SVM). Compared to the classifiers discussed previously, SVM is more suitable for studying non-linear process and capable to produce complex curved decision boundary between two classes in the training set but with higher risk of overfitting (Richard G. Brereton, 2009).The concept of SVM can be explained by the basic definition of simple linearly separable classes as illustrated in Figure 3 and subsequently extended to non-linearly separable case with the use of kernel functions and then the generalised case with penalty parameter (nonlinear machines trained on non-separable data) (Richard G. Brereton & Lloyd, 2010). By definition, training points lying on one of the hyperplanes are called support vectors (Burges, 1998; Xu, Zomer, & Brereton, 2006). Hence, the SVM model is built using support vectors which consist of small number of samples from the original dataset that lie near the decision boundary. However, in the case of overlapping classes, data points on the "wrong" side of the discriminant margin are weighted down to reduce their influence ("soft margin"). Thus, soft margin SVMs tolerate the classification error with the complexity of the model (Gunn, 1998). In nonlinearity case appropriate kernel function nonlinearly transforms samples or input space into a higher dimensional space (feature space) where the data points effectively become linearly separable. Consequently, the hyperplane in feature space is projected back into the input space to describe a non-linear boundary which gives optimal classification (Richard G. Brereton, 2009). There are four basic kernels commonly used that are Radial Basis Function (RBF),

Polynomial Function (PF), Linear Function (LF) and Sigmoid Function (SF) (Hsu, Chang, & Lin, 2003).In this work, we used soft margin SVMs with a RBF kernel which has two parameters to be optimised i.e. γ and σ. The value of these two parameters, i.e. γ is set to 1 and σ is set to the average of standard deviation of the all variables in the dataset to train the whole training set.

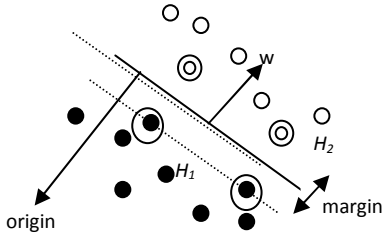


Fig. 3: Linear separating hyperplanes for the separable case. The support vectors are circled.

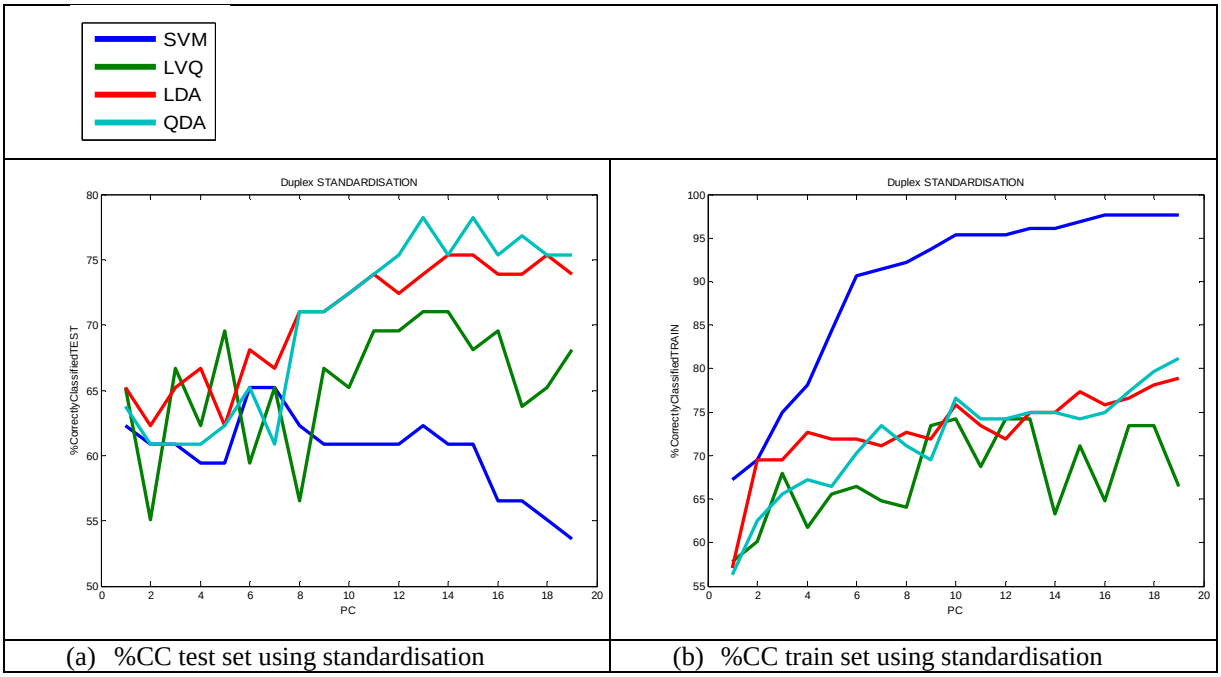
RESULT AND DISCUSSION

The model performance indicator of the classifiers and data splitting methods used in this work is percentage correctly classified (%CC) calculated on training and test set. The test set %CC gives a more realistic representation of the classification ability of the model. A well balanced model (i.e. one that does not suffer from over fitting) would have training and test set %CC that are similar. However, the %CC for the training set tends to be higher than that for the test set for models that are complex and prone to over fitting, (Dixon & Brereton, 2009). The models were developed using PCA scores as the variable reduction method using varying number of PCs ranging from 2 up to 20. Subsequently, the classification ability of each model was compared.

The results of classification using Duplex as data splitting method for both the training and test set are illustrated by graphs where the vertical axis represents the percentage correctly classified values while the horizontal axis correspond to increasing number of PC and each classification model is presented as different line colour. The values %CC of using Duplex method are moderately high for all classifiers that range approximately from 50% to 90% indicating good prediction with the training set results being marginally better.

It can be seen from the line graphs shown in Figure 4 that QDA classifier followed by LDA classifier marked as light blue and red line respectively perform reasonably well with high value of test set %CC. Based on the results of QDA classifier, the value of %CC for test set is highest at 14 PC (78%). Similarly, the maximum value for LDA is at 15 PC (71%). Hence, QDA is slightly better due to the distinct difference of variance-covariance matrices for the two classes. Moreover, the difference between %CC on training and test sets is relatively low which is about 2-3% for both classifiers. Hence, their performance are more stable and less vulnerable to overfitting.

Generally, all classifiers exhibit increasing value of percentage correctly classified training set with increasing number of PCs and the pattern is significantly obvious for SVM model. Furthermore, the difference of %CC between training and test s*et al*so increases considerably demonstrating higher risk of overfitting with increasing number of PCs for SVM model. For example, the difference is 5% at 2 PC and increases sharply to 33% at 10 PC as evident from Figure 4(a) and (b).On the other hand, LVQ model has the lowest value of percent correctly classified for both training and test sets compared to the other three models in all cases implying the poor performance of the model in prediction of artemisinin dataset as illustrated in Figure 4 as green line. After all, the recommended model for artemisinin dataset is QDA with standardisation as the data pre-processing method in Duplex data splitting approach.

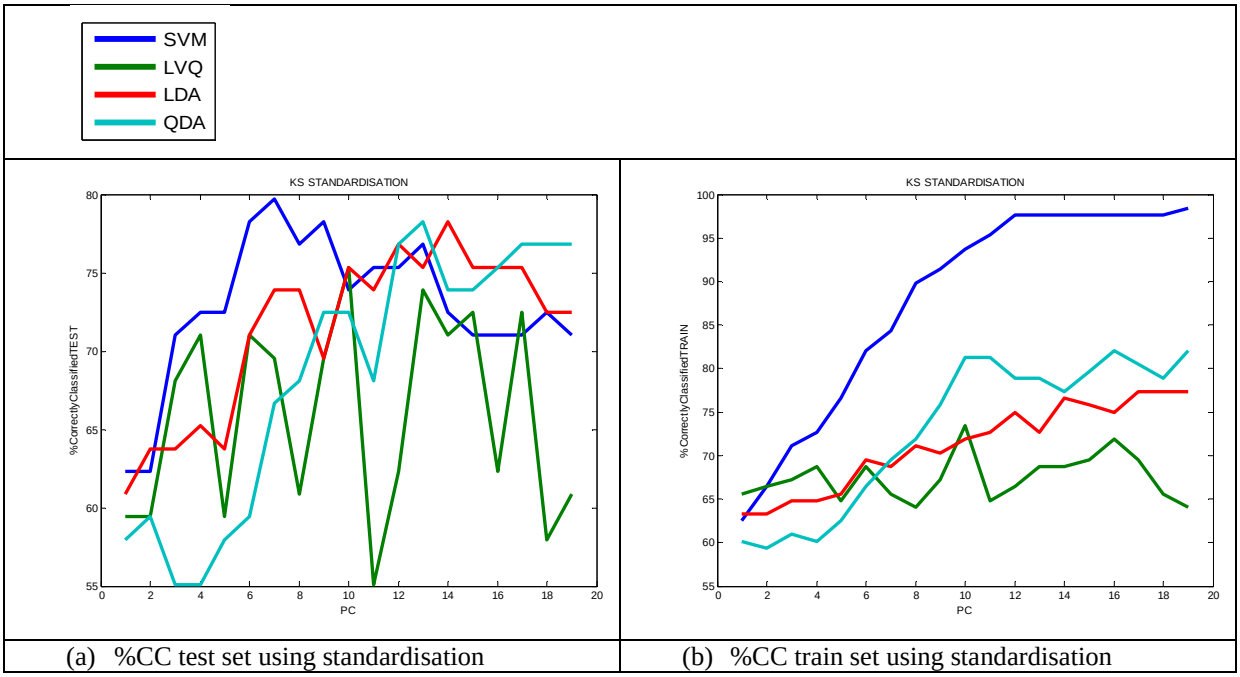


(a) %CC test set using standardisation (b) %CC train set using standardisation

Fig. 4: Comparison of the percentage correctly classified for SVM, LVQ, LDA and QDA for training and test set using Duplex method at 20 PC (line graph)

The results of applying Kennard-Stone or Kenstone as data splitting approach on artemisinin data set are displayed inFigure 5. Basically, all classifiers using Kennard-Stone method show good classification ability judging from the %CC for both training and test set values that are larger than 50%. In other words, all the classifiers predict more than 50% samples in the test set correctly.The trend seen in Duplex method is also seen in the Kennard-Stone method. The %CC training set values show increasing trend with increasing number of PC but the values are more significant in SVM model. SVM classifier appears to be overfitting the data with a visible difference between the training and test results. It is interesting to note that the %CC for LVQ is apparently lower as compared to the other models similar to Duplex method and hence least recommended for classification of artemisinin dataset.

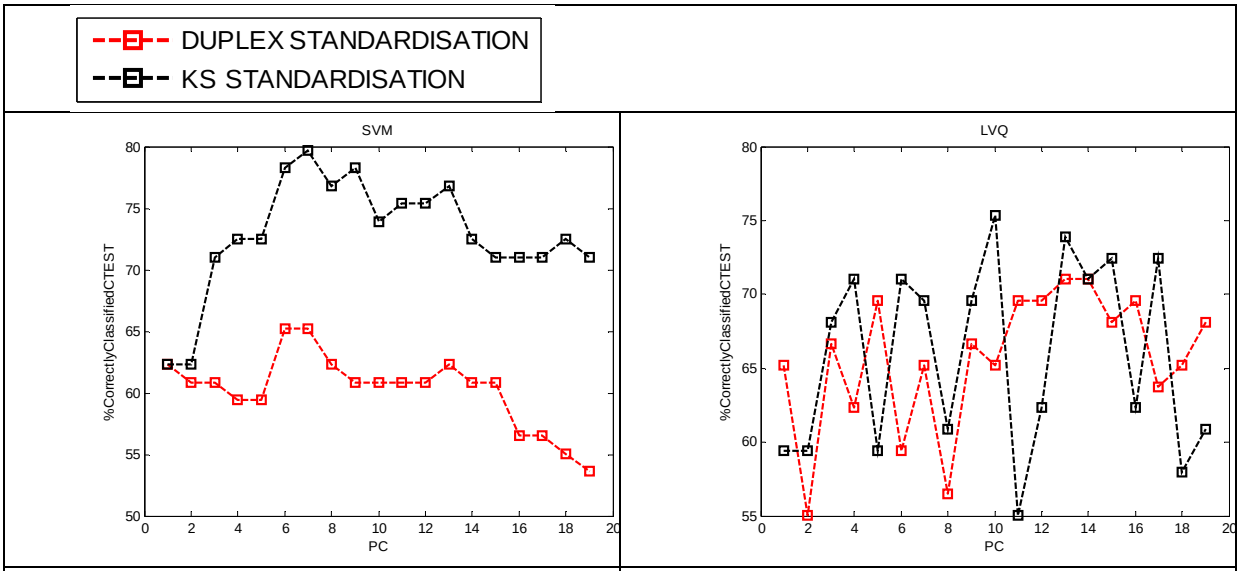
Referring to Figure 5, SVM model has the highest value of %CC of test set with maximum peak occurs at 8 PC (79%). Besides, the %CC of test set for LDA and QDA performs equally well and reaches their maximum values at 15 PC and 14 PC (78%) respectively. However, they have lower risk of overfitting with 1-2% difference between %CC training and test set as compared to SVM model.Although SVM model has its maximum peak at lower PC (8PC) with slightly higher percentage correctly classified of test set (79%), QDA and LDA models are much better since they have less risk of overfitting due to the very small difference between %CC training and test set. This might be due to the multi-normal data which gives advantage to QDA while SVMs and LVQ are prone to overfitting for data that are normally distributed. According to Dixon, the choice of appropriate method should also depend on the distribution of samples in variable sample (Dixon & Brereton, 2009).



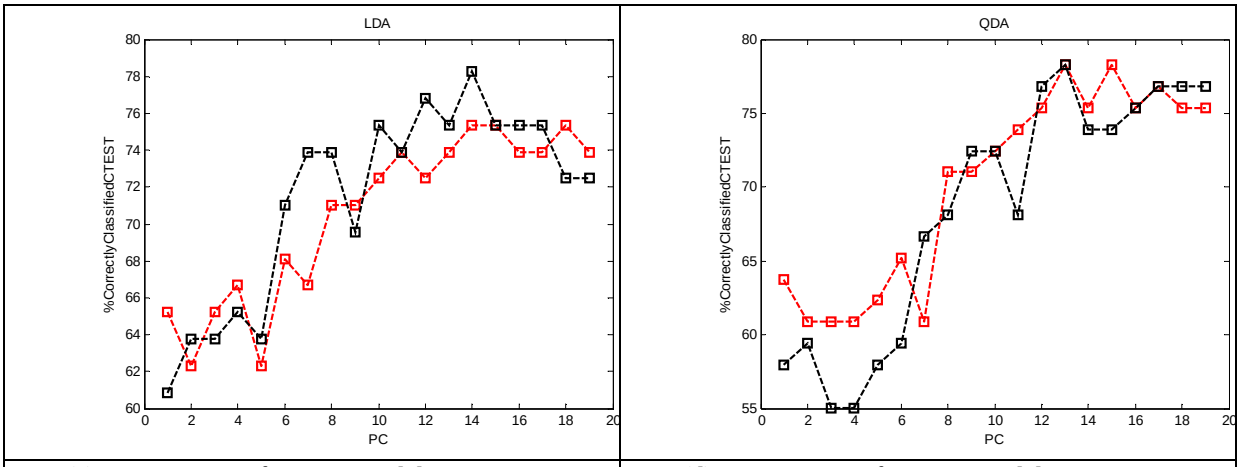
(a) %CC test set using standardisation (b) %CC train set using standardisation

Fig. 5: Comparison of the percentage correctly classified for SVM, LVQ, LDA and QDA for training and test set using Kennard-Stone method at 20 PC (line graph)

The graph in Figure 6 clearly shows that the results from Kenstone data splitting models indicated by the %CC of both test and training set for all classifiers are slightly higher than the models with the same PCs in Duplex method where the black lines representing Kennard-Stone are slightly above the red line in Figure 6.It can be seen from Figure 6 (a) that SVM classifier has good classification ability with higher percentage correctly classified for test set when using Kennard-Stone as data splitting method. However, the performance of the other three classifiers is approximately the same regardless of the method used. SVM classifier predicts better at lower PCs compared to LDA and QDA in Kennard-Stone but poorly in Duplex method. However, LDA and QDA models are better in predicting the test set with percentage value greater than 75% in both cases.



(a) %CC test set for SVM model (b) %CC test set for LVQ model



(c) %CC test set for LDA model (d) %CC test set for QDA model

Fig. 6: Comparison between Duplex and Kennard Stone of the percentage correctly classified of test set for the four classifiers using standardisation

Conclusion:

It can be summarized from the study that Kennard-Stone and Duplex algorithms ensure that selected samples are as representative as possible of the data set. However, this study showed that these two algorithms may not lead to models with similar performances. Hence, certain data splitting method is more suitable for certain classifier and data set. In this work, Kennard-Stone data splitting method and standardisation provide good combination in producing reliable predicted models. The difference between %CC training and test set is most striking in SVM results suggesting higher risk of overfitting.Generally, the performance of LDA model in Kennard-Stone data splitting is stable and the best compared to the other three classifiers.

It is suggested to perform other method of data pre-processing such as mean centering and row scaling. Besides that, variable selection can be carried out as an additional step in this work as well as using additional model evaluators such as percentage predictive ability and percentage models stability besides the percentage correctly classified test and training sets for future study. Moreover, it is recommended to repeat the division into training and test set 100 times or more using different selection of samples each time because the choice of samples included in the training set may influence the classification ability. Lastly, since both Duplex and Kennard-Stone are based on molecular descriptors alone, therefore other data splitting methods that are based on activity values would be recommended as comparison.

REFERENCES

Avery, M.A., M. Alvim-Gaston, C.R. Rodrigues, E.J. Barreiro, F.E. Cohen, Y.A. Sabnis, *et al*. 2002. Structure-Activity Relationships of the Antimalarial Agent Artemisinin. 6. The Development of Predictive In Vitro Potency Models Using CoMFA and HQSAR Methodologies. J. Med. Chem., 45: 292-303.

Brereton, R.G., 2006. Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data. TrAC Trends in Analytical Chemistry, 25(11): 1103-1111.

Brereton, R.G., 2009. *Chemometrics for Pattern Recognition*. Chichester, UK: John Wiley & Sons, Ltd.

Brereton, R.G., & G.R. Lloyd, 2010. Support Vector Machines for classification and regression. Analyst, 135: 230-267.

Burges, C.J.C., 1998. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2: 121-167.

Capron X., B. Walczak, O.E. Noord de, & M. D.L., 2005. Selection and weighting of samples in multivariate regression model updating. Chemometrics and Intelligent Laboratory Systems, 76: 205-214.

Cazelles, J., A. Robert, & B. Meunier, 2001. Alkylation of heme by artemisinin, an antimalarial drug. C. R. Acad. Sci. Paris, Chimie / Chemistry, (4): 85-89.

De Maesschalck, R., D. Jouan-Rimbaud, & D.L. Massart, 2000. The Mahalanobis distance. Chemometrics and Intelligent Laboratory Systems, 50(1): 1-18.

Dixon, S.J., & R.G. Brereton, 2009. Comparison of performance of five common classifiers represented as boundary methods: Euclidean Distance to Centroids, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Learning Vector Quantization and Support Vector Machines, as dependent on data structure. Chemometrics and Intelligent Laboratory Systems, 95(1): 1-17.

Ferreira, J.E.V., A.F. Figueiredo, J.P. Barbosa, M.G.G. Cristino, W.J.C. Macedo, O.P.P. Silva *et al*. 2010. A study of new antimalarial artemisinins through molecular modeling and multivariate analysis. Journal of the Serbian Chemical Society, 75(11): 1533-1548.

Guha, R., & P.C. Jurs, 2004. Development of QSAR Models To Predict and Interpret the Biological Activity of Artemisinin Analogues. J. Chem. Inf. Comput. Sci., 44(4): 1440-1449.

Gunn, S.R., 1998. Support Vector Machines for Classification and Regression. in Online reference manual.

Hsu, C.W., C.C. Chang & C.J. Lin, 2003. A Practical Guide to Support Vector Classication. in Online reference manual.

Jaafar, M.Z., 2011. *Chemometrics and Pattern Recognition Methods with Applications to Environmental and Quantitative Structure-Activity Relationship Studies.* University of Bristol.

Kamchonwongpaisan, S., & S.R. Meshnick, 1996. The Mode of Action of the Antimalarial Artemisinin and its Derivatives. Gen. Pharmac., 27(4): 587-592.

Lloyd, G.R., R.G. Brereton, R. Faria, & J.C. Duncan, 2007. Learning Vector Quantization for Multiclass Classification: Application to Characterization of Plastics. J. Chem. Inf. Model., 47(4): 1553-1563.

Meshnick, S.R., 2002. Artemisinin: mechanisms of action, resistance and toxicity. International Journal for Parasitology, 32: 1655-1660.

Olliaro, P.L., R.K. Haynes, B. Meunier, & Y. Yuthavong, 2001. Possible modes of action of the artemisinin-type compounds. TRENDS in Parasitology, 17(3): 122-126.

Ploypradith, P., 2004. Development of artemisinin and its structurally simplified trioxane derivatives as antimalarial drugs. Acta Tropica, 89(3): 329-342.

Posner, G.H., J.N. Cumming, P. Ploypradith, & C.H. Oh, 1995. Evidence for Fe(IV)=O in the Molecular Mechanism of Action of the Trioxane Antimalarial Artemisinin. J. Am. Chem. Soc., 117: 5885-5886.

Posner, G.H., S.B. Park, L.S. Gonza´lez, D. Wang, J.N. Cumming, D. Klinedinst, *et al*. 1996. Evidence for the Importance of High-Valent FedO and of a Diketone in the Molecular Mechanism of Action of Antimalarial Trioxane Analogs of Artemisinin. J. Am. Chem. Soc., 118: 3537-3538.

Puzyn, T., A. Mostrag-Szlichtyng, A. Gajewicz, M. Skrzyn & A.P. Worth, 2011. Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models. Struct Chem, 22: 795-804.

Todeschini, R., V. Consonni, A. Mauri, & M. Pavan, 2010. DRAGON - Software for Molecular Descriptor Calculations (Version 6.0 for Windows). Milan, Italy: Talete srl.

Xu, Y., S. Zomer, & R.G. Brereton, 2006. Support Vector Machines: A Recent Method for Classification in Chemometrics. Critical Reviews in Analytical Chemistry, 36(3-4): 177-188.