TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 11898 - 11907

Extensive Incorporation of K-Nearest Neighbor to Support Vector Machine through Correlation Studies for a Better Classification

Doreen Ying Ying Sim

Article Info
Volume 82
Page Number: 11898 - 11907
Publication Issue:
January-February 2020

Abstract
Correlation studies are performed to assess the patterns of data for all the applied datasets before k-Nearest Neighbor (kNN) algorithms are incorporated to the Support Vector Machine to form the algorithms to be developed. Based on the correlation studies on the data of the applied datasets, each of them is categorized into three categories of low, medium or high correlations. Weighted distances of kNN are then computed based on the correlation studies in order to tune and adjust the hinge loss function and width of the Support Vector Machine (SVM) kernel. The proposed formulations of SVM which are derived from the studies on the correlations among the patterns of data from the applied datasets are applied to tune the kernel width and therefore adjust the hinge loss function of SVM. When the adjusted SVM hinge loss function and its optimized kernel width after being computed by the proposed formulations, it is found that for all datasets, having low, moderate or high correlation coefficients, the proposed formulations and algorithms can optimally adjust the SVM Gaussian kernel width so that its hinge loss function can be tuned. After implementing the proposed and developed correlation-based k-Nearest Neighbor Support Vector Machine (ckNNSVM) algorithms to the datasets, it is shown to be more accurate in classification when compared with the classical SVM classification and the kNNSVM algorithms without getting the SVM hinge loss function tuned and/or the Gaussian kernel width adjusted accordingly through extensive correlation studies.

Keywords: *Correlation-based k-Nearest Neighbor Support Vector Machine (ckNNSVM) algorithms, Gaussian kernel width, k-Nearest Neighbor algorithms, support vector machine.*

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 21 February 2020*

I. INTRODUCTION

A. Classification Strengths of SVM and kNN Method

While support vector machine (SVM) is considered a sophisticated, robust and well-established classifier [4], [9]–[10], [12], [15], in terms of its classification strength even though data in the applied datasets are having weak relationships, k-Nearest Neighbor (kNN) on the other hand, is an instance-based learning or lazy learning method, which is non-parametric and is not considered a strong classifier when data in the dataset applied is not correlated well to each other. Therefore, this research is trying to study the correlation coefficients of the

data in the applied datasets before re-formulating the primal and dual problem classical formulations of SVM. Pearson correlation studies are performed for continuous variables in the applied datasets so as to assess the strength of the relationships among data when being correlated with each other. Based on the previous researches done on the datasets of Obstructive Sleep Apnea, i.e. OSA actual datasets*, as well as from the UCI online data repositories [3]–[8], [11], [13]–[14], the derived weighted distances by the kNN methods after the correlation

TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 11898 - 11907

studies have been performed are used to tune and monitor the hinge loss functions or the cost functions of SVM in this research. Both the SVM and kNN are popular machine learning algorithms [4], [9]–[10], [12], [15] and they are incorporated synergistically to improve the classification accuracies of the datasets being applied.

B. Research Background leading to the incorporation of SVM to kNN Method

In order to optimize the generalization performance of SVM and kNN, one of the most challenging problems of SVM and kNN is that when each of them is applied to the datasets which are having low correlation coefficients, i.e. due to having high data dimensionality (i.e. with number of data dimensions of more than 10) or too randomly distributed data, the auto default kernel width of SVM or the kNN methods will not achieve its optimal effects in terms of classification. In primal formulations of SVM, misclassifications are always allowed, but at the cost of a regularization parameter, i.e. β [4], [9]–[10], [12]. In dual formulations of SVM, hinge loss or cost function of SVM is used for maximum margin classification [4], [12], [15]. Classification strength of SVM decreases when regularization parameter cannot be optimally matched to control the trade-off function [1]–[2], [4], [9]–[10], [15]. The developed algorithms in this research are known as the correlation-based k-Nearest Neighbor Support Vector Machine (ckNNSVM) algorithms [4], [15].

By using correlation studies to incorporate kNN methods to SVM, datasets which have weak or moderate or strong relationships among data after the correlation studies will get the proposed formulations of this research to adjust the classical Gaussian kernel width of SVM. To solve the general problem background of SVM due to highly uncorrelated data and/or high dimensional data, correlation-based kNN methods use metrics to measure the similarity within patterns and Euclidean distance between patterns [4], [15]. Then, the proposed formulations of this research based on the metric derived from correlation-based kNN methods are used to tune and adjust the Gaussian kernel width of SVM. Classification accuracies will be shown to be significantly improved after applying the proposed new formulations of this research.

II. PRELIMINARY STUDIES AND RESEARCH BACKGROUND

A. Theoretical Background and Research Hypotheses

Based on the Structural Risk Minimization (SRM) of SVM, it has a Gaussian kernel width which when being used for classification or regression, is always indefinite in its numerical value although a default value of it is placed in the SVM setting, or know as the white-box of SVM [2], [4]. The cost parameter of SVM, i.e. β, as the penalty term which regulates all slack variables, ξ, of SVM (Equation (1) as shown), can control separation of hyper-plane in primal form [1]–[2], [4], [9]–[10], [12], [15].

$$\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + \beta \sum_{i=1}^{m} \xi_i \qquad (1)$$

$$\text{s. t.} \quad y_i(w^T X_i + b) \geq 1 - \xi_i \quad i = 1, \ldots, m, \ \xi_i \geq 0$$

where y_i is the sign of hyper-plane that is shown by equation of $(w^T X_i + b = 0)$, w is the input vector, and m is the maximum number of elements in the class assigned.

Equation (2) shows the above minimization problems of Equation (1) can be generally and holistically viewed as Structural Risk Minimization when SVM incorporates with other classifiers such as weak classifiers, i.e. decision trees, neural networks or the like in optimizing and maximizing the soft-margin SVM kernel functions [1]–[4]. In Eq.(2), 'R' stands for the structural 'risk', α stands for the weight factor such as the β in Eq.(1), and 'T' stands for the weak classifiers such as decision trees, neural networks or the like [13]–[14].

$$R_\alpha(T) = R(T) + \alpha |\tilde{T}| \qquad (2)$$

January - February 2020
ISSN: 0193 - 4120 Page No. 11898 - 11907

Research Hypothesis: This research hypothesizes that by figuring out the relationships among the training data or between the training data or training samples and testing data points before input to the typical SVM classification can optimally adjust the Gaussian RBF kernel width which can then minimize the trade-off of SVM. This has the same pursuit as modifying the hinge loss or cost function in the primal formulation of SVM (see Eq.(1)). In the same vein, Eq.(2) has the same pursuit in modifying the kernel width or function of SVM. Therefore, SVM classification accuracies can significantly be improved.

B. Research Question

The main research question to be researched upon is elaborated as below.

The RESEARCH QUESTION (RQ) is that to find out the correct proposed formulations derived from the correlation studies based on the distance and/or similarity among patterns, input in a weighted distance metric format by kNN methods, can tweak and work synergistically to match the SVM Gaussian kernel width parameter, σ. This is to ensure that datasets that have been pre-processed by the correlation studies on the datasets and by the kNN methods, the proposed and developed algorithms, i.e. ckNNSVM algorithms, can be optimally matched with SVM kernel width σ after being applied with the proposed formulations so that classification accuracy of each dataset being applied can significantly be improved.

C. Determination of the Pearson's Correlation and Correlation Studies

Throughout this paper, the determination on the correlation coefficients and correlation studies are based mainly on 3 criteria. These are elaborated as below.

(I) The FIRST CRITERIA (1) is based on the correlation coefficients that are obtained from the correlation studies applied to the datasets. Datasets which have low correlation coefficients, i.e. weak relationships among data, and in this case, a Pearson's correlation coefficient, denoted as 'κ', if κ<0.30, is considered as datasets with low correlations in the variables involved. Whereas, a correlation coefficient of κ>=0.30 AND κ<0.50, is considered as datasets with moderate correlations in the variables involved; and, correlation coefficient of κ>=0.50, is considered as datasets with high correlations in the variables involved. The symbol used for the correlation coefficient is denoted as the 'κ' sign. In Eq. (3) *add_corr*(x), or *aCr*, is denoted as the added value to the *q* value (refer to the 3rd CRITERIA for definitions on *aCr* values) on the adjusted Gaussian Kernel width of SVM based on the correlation studies on the relationships among data in each of the applied datasets.

$$add\_corr(x) = aCr = q + \frac{1}{8} Corr(x, x_i) = q + \frac{1}{8\kappa} \quad (3)$$

where x $\in$ input data,

x_i $\in$ data in the nearest neighborhood

(II) The SECOND CRITERIA (2) is the redefinition of the impact range of *aCr*, or added value to the *q* value, on the SVM kernel scale 'readjustment'. The level of impact of *aCr* on the readjustment of Gaussian kernel scale is divided into three(3) categories, i.e. heavy, normal and light. These three(3) levels are based on the impact of *aCr* to kernel scale ranges. Light impact of *aCr* to kernel scale, σ', produces an inclusive range of 0.01-4.00; normal impact of *aCr* to kernel scale, σ', produces an inclusive range of 4.01-8.00; then, heavy impact of *aCr* to kernel scale, σ', produces an inclusive range of 8.01 and above. Table 2 and Table 3 showed the different levels of impact of *aCr* to SVM Gaussian kernel width, i.e. σ'.

(III) The THIRD CRITERIA (3) is the redefinition of the range of *aCr* values for kernel scale readjustment. This criteria is based on the power value of *q* to readjust the SVM Gaussian new kernel width σ' after the *aCr* values have been determined by correlation studies and *k*NN has been applied to the ckNNSVM algorithms. Narrow readjustment of σ' by *aCr* can produce a range of 0.01-1.00; normal readjustment of σ' by *aCr* can produce a range of

January - February 2020
ISSN: 0193 - 4120 Page No. 11898 - 11907

1.01-3.00; and wide readjustment of σ' by *aCr* can produce a range of 3.01 and above.

III. PROPOSED FORMULATIONS FOR DATASETS OF DIFFERENT CORRELATIONS

A. Proposed Formulations and Approaches based on Correlation Studies

Equation (4) below indicates the reciprocal correlation between the σ or σ^2, i.e. kernel width and its function which in this case, the SVM RBF (Radial Basis Function) kernel function. In Equation (5), if the reciprocate of σ^2, or λ, is large, the kernel width is narrower and will fall off more rapidly when data points from training samples, i.e. X_s, move away to input vectors, i.e. X_i. If λ is reduced to a small value, so that the kernel width is wider, it will fall off slower but may need more computation time [1]–[4], [9]–[10]. Since kNN is applied to pre-process the data points, less computation time is needed, the values of lambda or λ computed will be significantly smaller than those without getting pre-processing done by kNN [1]–[4], [10].

$$K(X_i, X_s) = \exp\left(-\frac{\|X_i - X_s\|^2}{2\sigma^2}\right) \quad (4)$$

From Eq.(4) to Eq.(5), λ is in reciprocal relationship with the σ or σ^2 kernel width. Since lambda or λ will be smaller after the dimensionality reduction performed by the instance-based learning or lazy learning methods, i.e. kNN methods. From the reciprocates as stated, the adjusted kernel width σ' will be significantly wider than the initial or original kernel width, i.e. σ.

$$K(X_i, X_s) = \exp\left(-\lambda\|X_i - X_s\|^2\right) \quad (5)$$

In view of Eq.(4), the unadjusted width, i.e. σ, or adjusted width, i.e. σ', can adapt the distribution of features and characteristics in Euclidean space. In lieu of Eq.(4) and Eq.(5) which just show that σ is in reciprocal relationship with the dot product or the square of the Euclidean distance between the data points in the input vectors and those from training samples, Eq.(6) and Eq.(7) show the relationships between the weighted distance, *d*, amongst data point patterns and the adjusted SVM Gaussian kernel scale or width.

$$d = \langle \Phi(x) - \Phi(y) \rangle = \frac{2^{2q}}{\sigma'^2} = \frac{2^{2\left(aCr - \frac{1}{8\kappa}\right)}}{\sigma'^2} \quad (6)$$

$$\sqrt{d} = \|\Phi(x) - \Phi(y)\| = \sqrt{\langle \Phi(x) - \Phi(y) \rangle} = \frac{2^{2q/2}}{\sigma'^{2/2}} = \frac{2^{\left(aCr - \frac{1}{8\kappa}\right)}}{\sigma'} \quad (7)$$

Insert Eq.(7) into Eq.(4), then substituting σ with σ', we get Equation (8) shown as below.

$$K(X_i, X_s) = \exp\left(-\frac{\|X_i - X_s\|^2}{2\left(\frac{2^q}{\sqrt{d}}\right)^2}\right) \quad (8)$$

As shown in Eq.(3), the added value of correlations, *aCr*, or *add_corr*(x), is the *q* value plus one eighth of the correlation coefficient, κ. For the above, $\langle \Phi(x) - \Phi(y) \rangle$ in Eq.(6) and Eq.(7) are the dot products in weighted form, i.e. the square of the Euclidean distances in weighted form between data points and patterns. From Eq.(6) to Eq.(10), *aCr* is the derived correlation added value while *q* is the acquired weighted distance (as indicated above in CRITERIA (3)) before and after kNN methods during the implementation of ckNNSVM algorithms. The square root of *d*, $\|\Phi(x) - \Phi(y)\|$, is the weighted distance derived from *add_corr*(x), or *aCr*, as indicated in Equation (3).

B. Proposed Formulations for Adjusted Kernel Width of SVM

If no correlation studies are being performed, this research adopts just the kNNSVM algorithms by incorporating kNN to SVM, this research applies Equation (9) shown below as the equation for the

January - February 2020
ISSN: 0193 - 4120 Page No. 11898 - 11907

modified kernel width (in the form of Radial Basis Function or RBF).

$$\sigma' = \frac{2^q}{\sqrt{d}} \quad (9)$$

The values computed for *aCr* after the application of kNN method is based on the THIRD CRITERIA stated above on the difference of weighted distances before and after pre-processing done by kNN methods.

$$\sigma' = \frac{2^{\left(q+\frac{1}{8\kappa}\right)}}{\sqrt{d}} = \frac{2^{aCr}}{\sqrt{d}} \quad (10)$$

However, the latest proposed formulations in this research is based on correlation coefficients and values from correlation studies after the application of ckNNSVM algorithms are indicated in Equation (10) above. This is based on Equation (3) stated above that 1/8 of the correlation coefficients κ are added to the q values to form the *aCr* values to adjust the SVM Gaussian kernel widths. The added value of correlations, i.e. *aCr* in Eq.(10), is used as the power of 2 (plus the reciprocal of the correlation coefficient κ but one eighth (1/8) of its value, or $1/2^3$ of its value) in the proposed formulations for ckNNSVM algorithms in order to figure out the newly adjusted kernel width σ'.

IV. EXPERIMENTAL RESULTS AND RESEARCH FINDINGS

A. Classification Accuracies prior and post of proposed implementation of changes

In Table 1, from the averaged Pearson's correlation coefficients and correlation studies, the three datasets* of Obstructive Sleep Apnea (OSA) all showed weak positive relationships among data in the variables involved, i.e. small values of correlation coefficients, i.e. κ . The three OSA datasets* and 'Breast Cancer' dataset which have low correlation coefficients, much wider weighted distances are derived from the proposed formulations through kNN method being applied in ckNNSVM algorithms.

Table 1. Details of datasets, averaged Pearson's correlation coefficients and derived weighted distances before and after kNN methods on ckNNSVM algorithms.

	All the initial values before the application of the stated formulations					
	Datasets applied (from UCI online data repositories and three OSA actual datasets*)	Categories based on the correlation coefficients computed by Equation (3)	No. of tuples or rows	Averaged correlation coefficient, κ on involved variables, in each dataset	Weighted distances before kNN to the ckNNSVM algorithms	Weighted distances after kNN to the ckNNSVM algorithms
1	OSA dataset 1*	Low	400	0.27	1.01	2.48
2	OSA dataset 2*	Low	500	0.23	1.96	3.25
3	OSA dataset 3*	Low	450	0.29	1.98	3.59
4	Breast Cancer	Low	569	0.29	1.65	3.32
5	Diabetes	moderate	768	0.46	4.05	4.43
6	Australian Credit	high	690	0.61	2.82	3.01
7	Ionosphere	moderate	351	0.43	0.45	1.11
8	Heart Disease	high	270	0.54	3.75	4.05

TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 11898 - 11907

9	Liver Disorder	high	345	0.59	5.70	5.88
10	Titanic	moderate	1309	0.45	5.61	6.21

In Table 1, the averaged correlation coefficient, κ, is highlighted in red for datasets which are belonged to category of correlations of 'low'; but is highlighted in black for datasets which are belonged to category of correlations of 'high'. In Table 2, the column which shows *add_corr*(x) or *aCr*, that is computed based on Eq.(3). As shown, the *aCr*, power obtained for the new kernel width computations, and the acquired weighted distances, i.e. *d*, are derived from the proposed formulations.

Table 2. Initial and new kernel widths as well as power *q* and difference *d* after correlation and *k*NN application.

	Datasets and values acquired after applying proposed formulations					
	Datasets applied (Urvine online repositories and three actual OSA datasets*)	Initial kernel width σ (by default)	*q*, power obtained from the difference or *d*	*aCr* by applying correlation and *k*NN (refer to Eq.(3)) before ckNNSVM	Derived difference in weighted distances, *d*	New kernel width σ' (after proposed formulations)
1	OSA dataset 1*	1.92	3	3.46	1.47	9.08
2	OSA dataset 2*	1.73	3	3.54	1.29	10.24
3	OSA dataset 3*	1.99	3	3.43	1.61	8.49
4	Breast Cancer	2.11	3	3.43	1.67	8.34
5	Diabetes	1.72	1	1.27	0.38	3.91
6	Australian Credit	1.66	0	0.20	0.19	2.64
7	Ionosphere	0.98	2	2.29	0.66	6.02
8	Heart Disease	2.02	1	1.23	0.30	4.28
9	Liver Disorder	2.34	0	0.21	0.18	2.73
10	Titanic	1.00	2	2.28	0.60	6.27

Table 2 showed changes of the kernel widths after the changes of the power *q* and derived difference *d* to the *aCr* by applying correlation studies and *k*NN to the *ck*NNSVM algorithms based on the aforementioned datasets. In most of the datasets shown in Table 2, results indicated that *aCr* contributed to the much wider new Gaussian kernel widths, i.e. σ', after each of the power *q* and acquired weighted difference, *d*, is changed due to the *ck*NNSVM algorithms application. This especially applies for datasets with *aCr* values of 3.1 and above, i.e. those highlighted in blue. Only datasets of 'Australian Credit' and 'Liver Disorder' with *aCr* values <1.0, the derived values highlighted in red, have the newly developed 'medium kernel widths'. If the derived weighted distances among data points, *d*, are far from each other, medium or coarse Gaussian kernel widths, i.e. σ', will most likely be selected over the fine kernel widths σ' (which are usually the auto default values). In Table 2, the acquired weighted difference in distances, *d*, and the correlation coefficient acquired from the applied formulations in the Equation (10), i.e. *aCr*, or the power *q*, are input to compute the newly adjusted SVM kernel width σ'. As shown in Table 2 above, the newly adjusted kernel width σ' is much wider than the initial one, i.e. σ.

TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 11898 - 11907

B. Verification of the Experimental Results

Receiver Operating Characteristics (ROC) curves, as shown in Figures 1 and 2, there is a certain level of significant improvement in classification accuracies on all the datasets applied, and the level of significance is always measured by *p* values. From the plots of the ROC curves, a conclusion can be made that when added value to the *q* value, *add_corr(x)*, or *aCr*, i.e. *aCr* = 1.1 and above, increases, no matter these are kNNSVM or ckNNSVM algorithms, before or after ckNNSVM application to the datasets (see Table 1), wider kernel readjusted width can significantly improve the SVM classification. This again further illustrates that the proposed formulations based on *aCr* values do successfully work on the original kernel width σ sizes such that optimal kernel width, i.e. σ', matches with the applied *aCr* values.

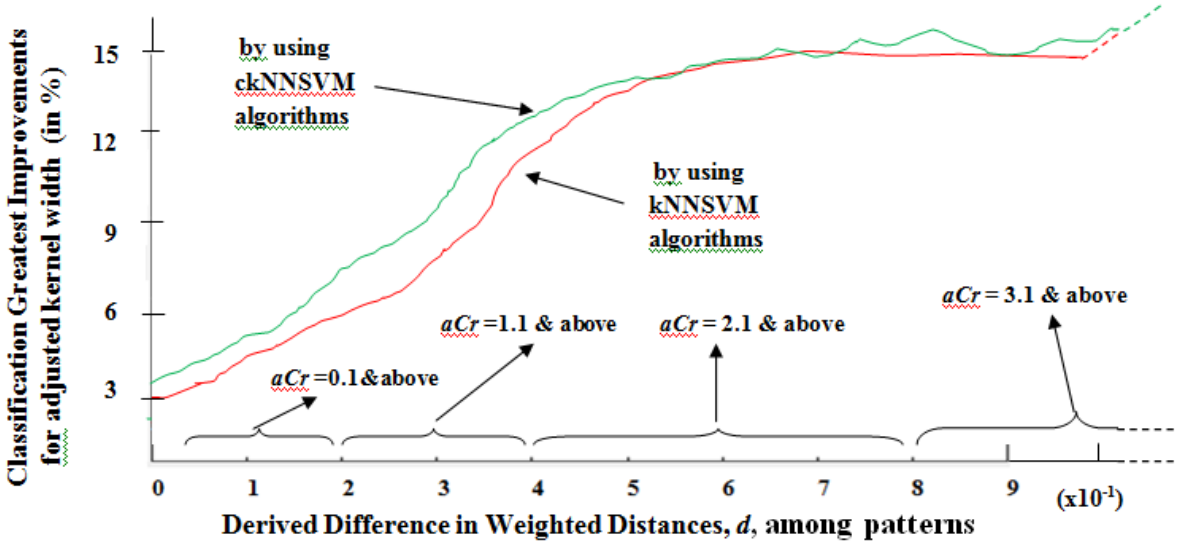

Fig. 1. Averaged ROC curves of all the datasets applied and each after ckNNSVM and kNNSVM algorithms application in achieving significant improvements in classification accuracies for adjusted kernel wdiths (in %) versus the acquired weighted difference in distances among data points (as shown in Table 2, in redefined scale of 0.1) at different readjusted kernel widths, i.e. σ', from the applied *aCr* values.

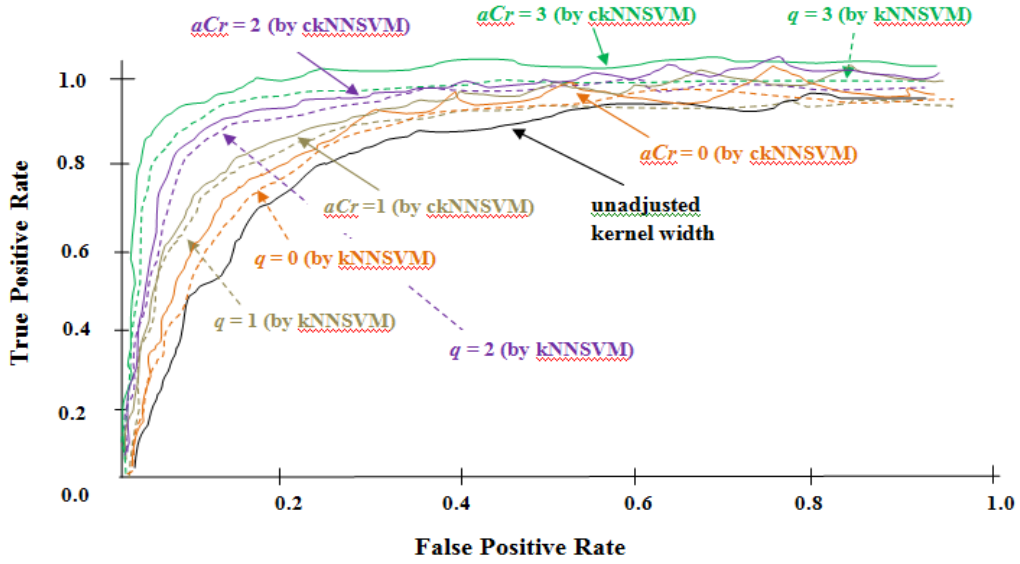

Fig. 2. Averaged generic outcome of the Receiver Operating Characteristics or ROC curves applied prior to as well as post ckNNSVM algorithms (σ' based on *aCr* values applied) as shown with solid lines on the stated datasets as compared with those dotted lines by kNNSVM algorithms (σ' based on *q*

January - February 2020
ISSN: 0193 - 4120 Page No. 11898 - 11907

values applied) where profound less improvements found for the q values as compared with those for the aCr values.

Figure 2 showed the generalized experimental outcome, i.e. ROC curves, prior to as well as post kNNSVM and ckNNSVM algorithms on the datasets stated. This is done through 4 different q values (by applying kNNSVM algorithms) and 4 different aCr values (by applying ckNNSVM algorithms) for the adjusted σ' kernel width, against those without getting its kernel width adjusted σ.

As shown in Table 3, the best classification accuracy (%) among the 3 different ranges (inclusive) of readjusted kernel widths σ' are highlighted in black. It shows a comparison between the classification accuracies and the greatest improvements (in %) achieved after the adjustment.

Table 3. Classification accuracies based on the different assigned kernel widths applied and the greatest improvements

	Datasets applied (from UCI repositories and OSA raw dataset)	Fine kernel width (by auto default) σ=0.01-1.00	Medium kernel width (adjusted) σ' =1.01-4.00	Coarse kernel width (adjusted) σ' = 4.01-12.00	Greatest improvements in classification accuracies (%)
		Classification Accuracy on each assigned σ' value of SVM			
1	OSA dataset 1*	67.9% [d]	67.5% [d]	76.0%	11.93[c]
2	OSA dataset 2*	73.2%	74.1%	82.2%	12.30[c]
3	OSA dataset 3*	77.5% [d]	77.7% [d]	85.8%	10.71[c]
4	Breast Cancer	68.4%	69.6%	72.4%	5.85[a]
5	Diabetes	75.3%	79.2%	84.0%	11.55[c]
6	Australian Credit	76.5%	83.3%	79.1%	8.89[b]
7	Ionosphere	67.6% [d]	67.9% [d]	72.9%	7.84[b]
8	Heart Disease	86.1%	87.2%	93.7%	9.01[b]
9	Liver Disorder	69.9%	75.6% [d]	75.4% [d]	8.15[b]
10	Titanic	88.7%	90.2%	96.0%	8.23[b]

[a]significance at $P<0.05$, [b]at $P<0.001$, [c]at $P<0.0001$; [d] about the same classification accuracies

V. CONCLUSION AND DISCUSSIONS

Experimental results by extensively incorporating k-Nearest Neighbor to SVM based on correlation studies have shown successful outcome for the proposed and developed algorithms in this research, i.e. ckNNSVM algorithms, as well as the proposed formulations, i.e. Equations (3) and (10), as per stated. Research question has been well answered and research hypotheses have been shown to be valid. The ckNNSVM algorithms have been shown to perform better than kNNSVM algorithms and also much better than the classical SVM classification without getting the kernel width sizes readjusted. Datasets of 'Australian Credit' and 'Liver Disorder', even though having the shortest weighted difference in distances

January - February 2020
ISSN: 0193 - 4120 Page No. 11898 - 11907

between data points, have shown significant improvements in the ckNNSVM classification accuracies with σ', i.e. the readjusted kernel widths.

By using the ckNNSVM algorithms, the correlation-based kNN methods that are applied to the data points which are not close to each other in terms of k-Nearest Neighbor weighted distances, can get the SVM kernel function adjusted so that its width is wider and drops off much more slowly. Therefore, in Tables 1 and 2, the four datasets having the lowest correlation coefficient of κ, i.e. three OSA datasets* and 'Breast Cancer' dataset, have the widest newly adjusted SVM Gaussian kernel widths. Shorter weighted difference in distances will need a finer or narrower adjusted kernel width , i.e. σ'.

Results showed that the acquired weighted difference in distances which are computed by kNN methods using the proposed formulations of this research, generic performance in terms of classification accuracies can significantly be improved even further. This is because the proposed wider or more optimized σ' will be more appropriate during the classification especially on datasets having low to medium correlation coefficients among data and patterns. The SVM auto default fine kernel width σ which its kernel function is much better to drop off more slowly rather than more rapidly when applied to datasets of low or medium correlations (but vice versa for datasets with high correlations such as 'Australian Credit' and 'Liver Disorder' datasets) as implemented in this research produces the least accurate classification accuracies. These are shown in the classification results in Table 3 under the 'fine kernel width' category, i.e. the 2nd column of Table 3. In short, the proposed and developed algorithms of this research, i.e. correlation-based kNNSVM algorithms, or ckNNSVM algorithms, can successfully apply the correlation studies on the datasets, to tune and tweak the kernel function and width such that correct kernel width and scale can be used to yield more accurate classification results. This betterment in classification accuracies applies to all types of datasets regardless of having low, medium or high correlation coefficients or values derived from correlation studies.

ACKNOWLEDGMENT

Datasets indicated with * sign are collected from the real patients'records of Obstructive Sleep Apnea (OSA) in the public hospitals and private Neurology clinics in Malaysia.

VI. REFERENCES

[1] S. Maldonado, J. Merigo, and J. Miranda, "Redefining support vector machines with the ordered weighted average," *Knowl. Based Syst.*, vol. 148, Mar. 2018, pp. 41–46.

[2] X. Gao, and J. Hou, " An improved SVM integrated GS-PCA fault diagnosis approach of Tennessee Eastman process," *Neurocomputing*, vol. 174, Jul. 2016, pp. 906–911.

[3] D. Y. Y. Sim, C. S. Teh, and A. I. Ismail, " Pushing constraints by rule-driven pruning techniques in non-uniform minimum support for predicting obstructive sleep apnea, " *Appl. Mech. Mater*, vol. 892, Jun. 2019, pp. 210–218.

[4] D. Y. Y. Sim, "Redefining the white-box of k-nearest neighbor support vector machine for better classification," *Computat. Sci. Tech.*, Jan. 2020, pp. 157–167, ACM Digital Library, Springer.

[5] D. Y. Y. Sim, C. S. Teh, and A. I. Ismail, " Pushing visualization effects into pushed schema enumerated tree-based support constraints, " *Appl. Mech. Mater.*, vol. 892, Jun. 2019, pp. 219–227.

[6] K. Shihab, D. Y. Y. Sim, and A. M. Shahi, "Angur: a visualization system for XML documents," May 2010, pp. 159–165 [*Dig. 9th WSEAS Int. Conf. Telecomm. & Informat.*, WSEAS, Catania, Italy, 2010, pp. 159–165].

[7] K. Shihab, and D. Y. Y. Sim, "Development of a visualization tool for XML documents," *Int. J. Comput.*, vol. 4(4), Aug. 2010, pp. 153–160.

[8] D. Y. Y. Sim, "Emerging convergences of HCI techniques for graphical scalable visualization: efficient filtration and location transparency of visual transformation," Jul. 2011, pp. 1–8 [*Dig. 7th Int. Conf. Info. Tech. in Asia*, Universiti Malaysia Sarawak (UNIMAS), Kuching, Malaysia, 2011, pp. 1–8].

[9] S. Maldonado, and J. Lopez, "Synchronized feature selection for support vector machines with twin hyperplanes," *Knowl. Based Syst.*, vol. 132, Dec.2017, pp. 119–128.

[10] B. D. Barkana, I. Saricicek, and B. Yildirim, "Performance analysis of descriptive statistical features in retinal vessel segmentation via Fuzzy Logic, ANN, SVM, and classifier fusion," *Knowl. Based Syst.*, vol. 118, Sep. 2017, pp. 165–176.

January - February 2020
ISSN: 0193 - 4120 Page No. 11898 - 11907

[11] D. Y. Y. Sim, C. S. Teh, and A. I. Ismail, "Adaptive apriori and weighted association rule mining on visual inspected variables for predicting Obstructive Sleep Apnea (OSA)," *Aust. J. Intell. Info. Process. Syst.*, vol. 14(2), Nov. 2014, pp. 39–45.

[12] X. Qiao, J. Bao, H. Zhang, F. Wan, and D. Li, "Underwater sea cucumber identification based on principal component analysis and support vector machine," *Measurement*, vol. 133, Jan. 2019, pp. 444–455.

[13] D. Y. Y. Sim, C. S. Teh, and A. I. Ismail, "Improved boosted decision tree algorithms by adaptive apriori and post-pruning for predicting obstructive sleep apnea," *Adv. Sci. Lett*., vol. 24, issue 3, Jan. 2018, pp. 1680-1685.

[14] D. Y. Y. Sim, C. S. Teh, and A. I. Ismail, "Improved boosting algorithms by pre-pruning and associative rule mining on decision trees for predicting obstructive sleep apnea," *Adv. Sci. Lett*., vol. 23, issue 11 , Feb. 2017, pp. 11593-11598.

[15] D. Y. Y. Sim, "Support vector machine pre-pruning approaches on decision trees for better classification," Sep. 2019, pp. 30–36 [*Dig. 2nd Int. Conf. Electronics & Electric. Eng. Tech.*, EEET2019, Penang, Malaysia, 2019, ACM New York, NY, USA, pp. 30–36].

AUTHORS PROFILE

Doreen Ying Ying Sim acquired her Doctor of Philosophy (PhD) in Computational Intelligence, Data Mining and Machine Learning after she graduated with MSc and BSc (Honors) degrees respectively in Business Information Technology and Computer Science respectively from University of Portsmouth and University of Central England, United Kingdom. She also has a double major Medical Sciences degree which she acquired from University of Otago, New Zealand. She has extensive lecturing and research experiences in Data Mining, Machine Learning and Computational Intelligence as well as certain research experiences in Artificial Intelligence. She has more than 14 recent publications, with certain research articles published in high impact factor ISI-indexed Tier-1 and SCOPUS-indexed international journals as well as other research papers in peer review international conference proceedings. Two of her publications were awarded with Best Paper Awards and another one of her publications was awarded with Best Presentation Award.