

# VC-based confidence and credibility for support vector machines

George E. Sakr · Imad H. Elhajj

Published online: 17 October 2014  
© Springer-Verlag Berlin Heidelberg 2014

**Abstract** Assigning a confidence and a credibility measures is a challenging stochastic inference problem. Some algorithms only yield the predicted value without evaluating the measure of confidence or credibility over the decision. Support vector machines (SVM) is one algorithm that showed state-of-the-art decision accuracy but lacks a measure of confidence and credibility over the decisions. In this paper we propose a new confidence measure based on the Vapnik and Chervonenkis (VC) dimension of a learning algorithm and the notion of complexity as defined by Kolmogorov. We also propose a new credibility measure based on the VC dimension. The resulting confidence and credibility measures are then tested on the well-known US postal handwritten digit recognition, on the Wisconsin breast cancer dataset and are also tested for agitation detection. The results show high and improved correlation between the decision and the confidence/credibility measures compared to Vovk's and Platt's methods.

**Keywords** Confidence · Credibility · Support vector machines · Digit recognition · Agitation detection

## 1 Introduction

The main drawback of the most machine learning algorithms resides in the lack of a confidence measure over the output prediction. For example most of the algorithms, if trained to

receive an image as an input and decide whether it represents the number “3” or number “8”, just output the decision: “3” or “8”. But they do not specify the decision is “3” with confidence of 80 %. And if the image happened to be very blurred, they do not indicate “lack of knowledge”. In 1983, Dawid distinguished between two different types of inferences: nominal inference and stochastic inference. Nominal inference is the decision itself while stochastic inference is a measure of the accuracy of the prediction (Dawid 1983). In the 1960s the idea of a universal randomness was developed by Kolmogorov (1968). His definition of randomness was used in Martin Löf's paper on defining random sequences, and he proved that indeed the definition of Kolmogorov does solve the problem of defining this type of random sequences. As it became clear later the notion of random sequence is closely connected to assigning confidence to predictions (Vovk et al. 1999). Unfortunately, Kolmogorov's algorithmic definition of randomness remained of purely mathematical interest for the simple reason that algorithmic measure of randomness is non-computable (Vovk et al. 1999). However, in his work, Kolmogorov suggested that randomness is a key application of complexity and that random sequences are irreducibly complex. In this paper, we exploit this relationship between VC dimension, complexity and randomness to propose a new continuous confidence measure. This measure is designed to be calculated for every decision. To the best of our knowledge, this is the first time that the VC dimension is used to define a confidence measure for SVM. We also propose a new VC-based continuous credibility measure that gives the ability to the learning algorithm to indicate: “lack of knowledge”. This is the case where the pattern that has to be classified is strange from what the algorithm is trained to do. In our definitions we do not try to find a computable approximation to Martin Löf's randomness tests. Instead, we use a measure of complexity provided by the VC dimension.

---

Communicated by V. Loia.

---

G. Sakr (✉) · I. Elhajj  
American University of Beirut, Beirut, Lebanon  
e-mail: ges07@aub.edu.lb

I. Elhajj  
e-mail: ie05@aub.edu.lb

The proposed measures are computed for every sample separately, differentiating them from the probability of error of a learning algorithm which is the average error over all samples. The rest of the paper is organized as follows: Sect. 2 presents a concise introduction of the theory of support vector machines; this section also highlights the limitations of SVM regarding its confidence and credibility measures output; Sect. 3 discusses previous techniques of defining confidence and credibility for different learning algorithms; Sects. 4 and 5 present the definitions of confidence and credibility and their theoretical framework. Finally, Sect. 6 shows the application of these measures on well-known datasets.

## 2 Support vector machines

The problem tackled in this paper is the binary pattern recognition problem. Having  $n$  labeled data vectors  $(x_1, y_1), \dots, (x_n, y_n)$ , where  $x \in \mathbb{R}^k$ ,  $y \in \{-1, +1\}$  and  $n$  is the number of labeled data vectors. The requirement is to classify an unknown new data vector  $x_{n+1}$ . Only one assumption is made about the data points  $x$ : all the examples are generated by a fixed but unknown stochastic process. This means that the generating mechanism has unknown but fixed probability distribution; hence all the data pairs  $(x_i, y_i)$  are independent and identically distributed (iid).

### 2.1 Theoretical background

In the simplest form, SVM uses a linear hyperplane to create a classifier with a maximal margin (Kecman 2001). In cases where the data are not linearly separable, the data are mapped into a higher dimensional feature space. This task is achieved using various non-linear mapping functions: polynomial, sigmoid and radial basis functions (RBF) such as Gaussian RBF. In the higher dimension feature space the SVM algorithm separates the data using a linear hyperplane. Not like other techniques, probability model and probability density functions do not need to be known apriori. This is very important for generalization purposes, as in practical situations there is not enough information about the underlying probability laws and distributions between the inputs and the outputs.

In the case of linearly separable data, the approach is to find among all the separating hyperplanes the one that maximizes the margin. Clearly, any other hyperplane will have a greater expected risk than this hyperplane. During the learning stage, the classifier uses the training data to find the parameters  $\mathbf{w} = [w_1 w_2 \dots w_n]^T$  and  $b$  of a decision function  $d(\mathbf{x}, \mathbf{w}, b)$  given by:

$$d(\mathbf{x}, \mathbf{w}, b) = \mathbf{w}^T \mathbf{x} + b = \sum_{i=1}^n w_i x_i + b \tag{1}$$

The separating hyperplane follows the equation  $d(\mathbf{x}, \mathbf{w}, b) = 0$ . In the testing phase, an unseen vector  $x$ , will produce an output  $y$  according to the following indicator function:

$$y = \text{sign}(d(\mathbf{x}, \mathbf{w}, b)) \tag{2}$$

In other words the decision rule is the following: if  $d(\mathbf{x}, \mathbf{w}, b) \geq 0$  then  $x$  belongs to class (1) and if  $d(\mathbf{x}, \mathbf{w}, b) < 0$  then  $x$  belongs to class (-1).

The weight vector and the bias are obtained by minimizing the following equation:

$$L_d(\alpha) = 0.5\alpha^T H \alpha - f^T \alpha \tag{3}$$

subject to the following constraints:

$$y^T \alpha = 0,$$

where  $H$  denotes the Hessian matrix given by  $H = y_i y_j (x_i x_j)$  and  $f$  is the unity vector  $f = [1, 1 \dots 1]^T$ . Having the solutions  $\alpha_{0i}$  of the dual optimization problem will be sufficient to determine the weight vector and the bias using the following equations:

$$\mathbf{w} = \sum_{i=1}^l \alpha_{0i} y_i x_i \tag{4}$$

and the bias is given by

$$b = \frac{1}{l} \sum_{i=1}^l (y_i - x_i^T \mathbf{w}), \tag{5}$$

where  $l$  represents the number of support vectors. The linear classifier presented above has limited capabilities since it is only used with linearly separable data while in most practical applications data are random and are not linearly separable. The non-linear data have to be mapped to a new feature space of higher dimension using a suitable mapping function,  $\Phi(x)$ , which is of very high dimension potentially infinite. Fortunately, in all the equations, this function appears only in the form of a dot product.

From the theory of reproducing kernel Hilbert spaces (Aronszajn 1950), which is beyond the scope of this paper, a kernel function is defined to be

$$K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle. \tag{6}$$

By replacing the dot product  $x_i \cdot x_j$  by  $K(x_i, x_j)$  in all the previous equations (3) becomes

$$L_d(\alpha) = \frac{1}{2} \sum_{i,j=1}^l y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{i=1}^l \alpha_i \tag{7}$$

subject to

$$\sum_{i=1}^l \alpha_i y_i = 0,$$

The decision hyper surface in Eq. (1) is given in the non-linear space by the following equation:

$$d(\mathbf{x}) = \sum_{i=1}^l y_i \alpha_i K(x_i, \mathbf{x}) \quad (8)$$

This remarkable characteristic of the kernel transformation gives the ability for support vector machines to operate on multi-dimensional data without affecting the processing time. Indeed in the linear case, the processing time is roughly the time needed to invert the Hessian matrix which is of  $\mathcal{O}(n^3)$ , where  $n$  is the number of training points. Since the transformation from the linear to the non-linear case is performed by the simple kernel transformation, the dimension of the Hessian matrix is not changed and hence the processing time is the same, thus its applicability and high performance in multi-dimensional data.

The solution of Eq. (3) yields the hard margin classifier. In general it is useful to use a soft margin classifier to preserve the smoothness of the hyperplane and prevent  $\alpha_i$  from tending to infinity. This classifier is obtained using the same minimization process by just adding one more constraint to Eq. (3). The constraint is  $0 \leq \alpha_i \leq C$ , where  $C$  is defined by the user. If  $C$  tends to infinity, the soft margin classifier tends towards the hard margin.

## 2.2 Limitation

Support vector machines provide state-of-the-art results in many applications for nominal inference. But unfortunately it has neither indication of the confidence nor the credibility over a certain prediction. In the literature, there exists several theorems that are used for measuring the performance of the algorithm (which is usually the probability of error), but they are in practice very challenging to implement. For example, Vapnik has introduced Theorem 5.2 in (Vapnik 1995) which, as stated in the introduction, is impossible to implement in practice because it involves computing an expected value based on the probability distribution of the training set, which is assumed to be unknown. Although this expectation can be estimated from only the training set, it will not give a tight bound because the estimation is very crude. Bayesian estimation has theoretically strong bounds on the probability of errors, but as it is known the Bayesian approach needs to have expressions of the underlying probability distribution of the training data. If the estimated distributions do not reflect the true distribution, then the bounds will not hold (Melluish

et al. 2001). In what follows, we discuss previous techniques to define confidence and credibility measures.

## 3 Related work

Support vector machines have been gaining grounds on other classification algorithms and have been used in many applications. Ercan used SVM to classify patents if they have good value or not (Ercan et al. 2014), while Liu et al. (2013) used SVM for scene classification of humanoid robot. However, it is very rare to find some work that defines a confidence or a credibility measure on SVM's decision. Vovk et al. used support vectors (SV) to implement an approximation for the notion of randomness. They relied on the fact that in support vector machines a training point that has a high Lagrange multiplier is a difficult point to classify; hence if  $(x_1, y_1) \dots (x_l, y_l)$  are training points and  $x_{l+1}$  is the testing point, then the machine is trained two times: the first is by giving the testing point the label  $-1$  and the second is by giving the testing point the label  $+1$ . If  $\alpha_i$  represents the Lagrange multiplier for the training point  $x_i$  then their definition of confidence follows these steps: first define the  $p$  value for that testing point given by

$$p \text{ value} = \frac{\#\{i : \alpha_i \geq \alpha_{l+1}\}}{l + 1}$$

The  $p$  value represents the fraction of the number of Lagrange multipliers that are greater than the multiplier of the testing point. If this number is large, then the multiplier corresponding to the testing point is small and hence the point is not difficult to classify. The  $p$  value is computed under both assumptions ( $-1$  and  $+1$ ) and the decision goes with the assumption that has the higher  $p$  value. The confidence is defined as  $1 - P_2$ , where  $P_2$  is the  $p$  value of the class that was not chosen, while the credibility is given by  $P_1$  which is the  $p$  value of the chosen class. The limitation of this method is that it discards the original decision of the support vector and replaces it with a new decision mechanism that has no reported error probability. Support vectors method was also used by Vapnik et al. (1998) where they based their study on the error bound defined in Vapnik (1995) (Theorem 5.2) that states that the error probability of the new testing point is at most:

$$\frac{E(\text{Number of support vectors among } x_1 \dots x_{l+1})}{l + 1}, \quad (9)$$

where  $x_1 \dots x_{l+1}$  are generated according to the fixed probability density  $P$  which is unknown. Obviously since  $P$  is unknown, it is impossible to compute Eq. (9). An approximation of its value is by training the SV two times: the first is by giving the testing point the label  $-1$  and the second is by giving the testing point the label  $+1$ . From Eq. (9) it is

clear that the error probability is increasing with the number of support vectors. So if the testing point is a support vector in one class and not in the other, then it will get classified in the class where it was not a support vector. And if it was a support vector in both the classes, it will get classified in the class that has the least number of support vectors. They also proved that the testing point will at least be a support vector in one of the two classes. The confidence over the decision is defined as:

$$1 - \frac{\#SV(\text{other class})}{l + 1}$$

The main limitation of this method is that it relies on the fact that the number of support vectors is very limited in comparison with the number of training points and also the definition of a new algorithm that does not take into consideration the strength of the decision of SVM, while providing no error bound on this new decision scheme. Confidence over machine learning algorithm decision was introduced in a slightly different framework by Platt (1999). In his work Platt introduced the notion of probability over the decision for SVM. After the training phase, the outputs of the training points were recorded. The log sigmoid function was used to fit the probability distribution of these outputs. For every new point corresponds an output that is mapped by the log sigmoid function into a probability which he claims to be the confidence over that decision. Platt stated that if a point has an output that is close to the output of the training points, then it must have a high confidence. The log sigmoid function used by Platt represents the posterior probability and is given by:

$$P(y = 1/f) = \frac{1}{1 + \exp(Af + B)},$$

where  $f$  represents the SVM output for the corresponding point and  $A$  and  $B$  are the parameters to be fitted using maximum likelihood estimation for the training set. The limitation of this method is the problem of overfitting the sigmoid function. If no cross-validation techniques are used, the sigmoid function will definitely face an overfitting issue. Even when using cross-validation Dumais (1998) and Joachims (1998) showed that the sigmoid can still face overfitting. Another limitation is the claim that if a point looks like a training point (the value of the decision is highly probable), it does not necessarily mean that the decision over that point has high confidence. Confidence measure has also been used in the development of the active learning algorithms. Active learning tries to resolve the problem of choosing the training points that has the most information. Li introduced a confidence measure based on the output of the support vector machines but without using a sigmoid function to model the output as the one used by Platt in (1999). A dynamic bin width allocation method was proposed to estimate the sample

conditional error for each data point; then the ones that have the highest probability of errors are used as training points because these points are the ones that have the highest information (Li and Chen 2006). Another confidence measure is introduced by Mitra et al. (2004) that is based on the  $k$  nearest neighbor algorithm and is combined with the distance to the hyperplane and integrated with support vector machines to also decide which points to use for training. Another confidence approach was done by our team (Sakr et al. 2010). This approach introduced a binary confidence measure using VC dimension for binary classification. The original training set is used to train a classifier "O". This classifier takes the decision on whether a testing point "x" is of class (-1) or class (+1). "x" is then added to the original training set first as class (-1) which is used to train a new classifier "M"; then as class (+1) which trains another classifier "P", the VC dimension of both classifier "M" and "P" was approximated. If the decision of the "O" coincided with the classifier that has the lowest VC between "M" and "P", then the decision is confident; otherwise, the decision is not confident. The drawback of this method is that this confidence was only a binary confidence measure and is not continuous between zero and one. In this paper, we propose a new continuous confidence measure and a new continuous credibility measure based on the VC dimension; we also present a study on the reference value and limiting values of these measures and the geometric interpretation of these values. In the next section, the proposed confidence and credibility measures are introduced and applied with support vector machines on the US postal dataset, the Wisconsin breast cancer dataset and agitation detection dataset.

## 4 Proposed confidence and credibility measures

Our proposed measure of strangeness or non-typicality is based on the VC bound introduced by Vapnik (1998). The next subsection is a concise introduction to the VC theory and to the bound that is used as a measure of strangeness.

### 4.1 VC theory and error bound

The proposed method is based on a dimension proposed by Vapnik and Chervonenkis which was named after them: the VC dimension. By definition, the VC dimension is the capacity of the learning algorithm to shatter points in the input space (Vapnik 1998). Formally it is the cardinality of the largest set of points that an algorithm can shatter. The importance of the VC dimension is that it appears explicitly in the bound of the total error of an algorithm. The total error of the learning machine is the sum of the training error (empirical error) and the testing error (generalization error):

$$\varepsilon = \varepsilon_{emp} + \varepsilon_g, \tag{10}$$

where  $\varepsilon_{emp}$  is the training error, and  $\varepsilon_g$  is the generalization error (Vapnik 1998).  $\varepsilon_{emp}$  can be made arbitrarily small by choosing a machine with a VC dimension at least equal to the number of training points. To decrease the training error, one has to increase the VC dimension. If a machine can split all the training points without any errors, it will have  $\varepsilon_{emp} = 0$ . Vapnik has established a bound on the testing error given by:

$$\varepsilon_g < \sqrt{\frac{VC \left[ \ln \left( \frac{2l}{VC} \right) + 1 - \ln \left( \frac{\eta}{4} \right) \right]}{l}}, \tag{11}$$

where  $l$  is the number of training data, VC is the VC dimension (referred to as  $h$  in some references), and  $1 - \eta$  is the probability for which this last equation holds. This inequality shows that the error is bounded by an increasing function of the VC dimension, and thus a trade-off should be made between the empirical error and the generalization error. Although it is extremely difficult and sometimes impossible to compute the VC dimension of a certain algorithm, a bound on the VC dimension has been established and will be very useful in building the confidence and credibility measures. Vapnik states that a bound on the VC dimension is given by

$$VC < VC_{max} = \|w\|^2 D^2, \tag{12}$$

where  $D$  is the minimum radius of the sphere that contains all the training points and  $\|w\|$  is the norm of the weight vector that SVM is minimizing.  $\|w\|^2$  is given by Vapnik (1998):

$$\|w\|^2 = \sum_{i=1}^n \alpha_i \tag{13}$$

This bound is important in two ways: it is easy to compute and Burges has shown that the true VC is closely related to this bound. In particular he showed that, in most of the cases, the true minimum of the VC dimension is obtained when this bound is minimal (Burges 1998). Hence in the rest of this paper the study is concentrated on  $VC_{max}$ , and for clarity purposes, we will omit the subscript and denote it by VC.

#### 4.2 Confidence and credibility definitions

Consider the set of training points  $(x_1, y_1), \dots, (x_n, y_n)$  and consider the testing point  $x_{n+1}$ . The classification  $d(x_{n+1})$  of  $x_{n+1}$  is given by the sign of Eq. (8). What is missing is a confidence measure over  $d(x_{n+1})$ . To define the confidence measure,  $x_{n+1}$  is considered as a training point of class  $-1$ , then as class  $+1$ . The training of the support vector is carried

out using the old training set, to which  $x_{n+1}$  is appended first as  $-1$ . The optimization problem yields the Lagrange multipliers vector  $\alpha_{-1}$  that corresponds to  $x_{n+1}$  being trained as class  $-1$ . Then the optimization is done using  $x_{n+1}$  labeled as  $+1$ . The optimization problem yields the vector of Lagrange multipliers  $\alpha_{+1}$ . The aim is to find the VC bound given by Eq. (12) for both cases ( $-1$  and  $+1$ ). Having both Lagrange multiplier vectors it is possible to find  $\|w\|$  by Eq. (13). It remains to find  $D$ . Since  $D$  is the radius of the smallest sphere englobing all training points, it is independent of the classification ( $-1$  or  $+1$ ). This problem was partially solved by Vapnik (1998) and in more detail by Scholkopf and Smola in (2002). They established that when the training points  $(x_1, y_1), \dots, (x_n, y_n)$  are mapped by SVM to a higher dimension space by the mapping function  $\phi$ , the center  $O$  of the sphere that englobes all training points is given by:

$$O = \sum_{i=1}^n \lambda_i \phi(x_i), \tag{14}$$

where  $\lambda_i$  is given by the following quadratic minimization equation:

$$L = \sum_{i=1}^n \lambda_i - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j K(x_i, x_j) \tag{15}$$

subject to the constraints:

$$\begin{aligned} \sum_{i=1}^n \lambda_i &= 1 \\ \lambda_i &> 0 \end{aligned}$$

“O” is the center of the sphere of smallest radius that contains all training points; thus it is possible to deduce the radius of the sphere by noting that the radius is the largest distance between “O” and any training point  $x_k$ :

$$\begin{aligned} D^2 &= \max_{x_k} \|O - \phi(x_k)\|^2 \\ &= \max_{x_k} \left\{ \left( \sum_{i=1}^n \lambda_i \phi(x_i) - \phi(x_k) \right) \left( \sum_{i=1}^n \lambda_i \phi(x_i) - \phi(x_k) \right) \right\} \\ &= \max_{x_k} \left\{ \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j K(x_i, x_j) \right. \\ &\quad \left. - 2 \sum_{i=1}^n \lambda_i K(x_i, x_k) + K(x_k, x_k) \right\} \tag{16} \end{aligned}$$

By iterating through the training points the radius  $D$  can be determined and hence, the VC bound given in Eq. (12) is computed for each case. In what follows,  $VC_{-1}$  will denote

the value of the bound when  $x_{n+1}$  is considered of class  $-1$ ,  $VC_{+1}$  when it is considered as class  $+1$  and  $VC_0$  is before adding  $x_{n+1}$  to the training set. The standard for a measure of confidence is to be a function  $C_f : \mathbb{R} \rightarrow [0, 1]$  where  $0$  is achieved for the point with the lowest confidence, and  $1$  is achieved for the point with the highest confidence. From Eq. (11) it is possible to make the following observations:

1. If the decision is  $-1$ , then  $VC_{-1}$  should be less than  $VC_{+1}$ ; moreover, the confidence should increase as  $VC_{+1}$  is further away from  $VC_{-1}$ .
2. If the decision is  $+1$ , then  $VC_{+1}$  should be less than  $VC_{-1}$ ; moreover, the confidence should increase as  $VC_{-1}$  is further away from  $VC_{+1}$ .

The observations are important to define a good criterion that captures both observations in one equation.

**Definition 1** The criterion over which the confidence is based is defined by

$$\zeta = \text{sign}[d(z)](VC_{-1} - VC_{+1}) \tag{17}$$

This criterion groups the analysis made earlier in a single equation. Indeed if condition 1 is true, then the confidence will be an increasing function of  $\zeta$  because  $\text{sign}[d(z)] = -1$ ; hence if  $VC_{-1} \ll VC_{+1}$  then  $\zeta \gg 0$ , and if condition 2 is true then the confidence is also an increasing function of  $\zeta$  because  $\text{sign}[d(z)] = 1$  and if  $VC_{-1} \gg VC_{+1}$  then  $\zeta \gg 0$ . Thus what remains is to define the function  $C_f$ :

$$C_f : \mathbb{R} \rightarrow [0, 1] \\ \zeta \rightarrow C_f(\zeta)$$

which is defined as:

$$C_f(\zeta) = \frac{1}{1 + \exp(-\zeta)} \tag{18}$$

This function is always positive because the numerator and denominator are positive and is always less than 1 because the denominator is always greater than the numerator. This function is an increasing function of  $\zeta$ , indeed as  $\zeta$  increases the exponential decreases and hence the function increases.

The standard for a measure of credibility is to be a function  $C_r : \mathbb{R} \rightarrow [0, 1]$  where  $0$  is achieved for the point with the lowest credibility, and  $1$  is achieved for the point with the highest credibility. From Eq. (11) it is possible to make the following observations:

1. If the decision  $d(x_{n+1}) = -1$  then, to be credible,  $VC_{-1}$  should be close to  $VC_0$  and possibly less. Indeed if  $VC_{-1}$  is much larger than  $VC_0$  then it is possible to deduce that the testing point is very far from the training points and

has induced a considerable change in the shape of the decision hyperplane. Thus the point is not credible. Otherwise,  $VC_{-1}$  can be close to  $VC_0$  and hence it is close to the training points and did not change the hyperplane considerably, thus the point is credible. Moreover  $VC_{-1}$  can be even less than  $VC_0$ . This shows that this point has made the hyperplane even better and must also be highly credible.

2. If the decision  $d(x_{n+1}) = +1$ , the same analysis as above is also valid by replacing  $VC_{-1}$  by  $VC_{+1}$ .

**Definition 2** The criterion over which the credibility is based is defined by:

$$\xi = (VC(d) - VC_0) \tag{19}$$

Where  $VC(d)$  is defined by:

$$VC(d) = \begin{cases} VC_{-1} & \text{if } d(x_{n+1}) = -1 \\ VC_{+1} & \text{if } d(x_{n+1}) = +1 \end{cases}$$

This criterion groups the analysis made earlier in a single equation. Indeed if  $d(x_{n+1}) = -1$ , then the credibility will be a decreasing function of  $\xi$ : if  $VC_{-1} \gg VC_0$  then  $\xi \gg 0$ , and if  $d(x_{n+1}) = 1$  then the confidence is also a decreasing function of  $\xi$  because if  $VC_{+1} \gg VC_0$  then  $\xi \gg 0$ . Thus what remains is to define the function  $C_r$ :

$$C_r : \mathbb{R} \rightarrow [0, 1] \\ \xi \rightarrow C_r(\xi)$$

which is defined as:

$$C_r(\xi) = \frac{1}{1 + \text{poslin}(\xi)} \tag{20}$$

where  $\text{poslin}(x)$  is defined by

$$\text{poslin}(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases}$$

The credibility function is always a positive because the numerator and denominator are positive and is always less than 1 because the denominator is always greater than or equal to the numerator. This function is a decreasing function of  $\xi$ ; indeed as  $\xi$  increases the denominator increases and hence the function decreases. This function yields a credibility of 1 if  $VC(d) \leq VC_0$ .

Once the confidence and credibility over the testing point have been computed, one question remains to be answered: what are the values that are considered as reference for confidence and credibility? In other words it is obvious that by considering two testing points:  $t_1$  and  $t_2$ , if  $C_f(t_1) > C_f(t_2)$

then the decision over  $t_1$  is more confident than the one over  $t_2$ . But is this analysis valid for the credibility? And is a confidence or credibility of 0.5, good or bad? In the next section we will define good reference points for confidence and credibility and we will study their properties as well as present practical methods to define them.

### 5 Limits of confidence and credibility

The analysis presented in this section is the same for both confidence and credibility. The analysis is made on confidence. Consider an SVM, trained with a defined set of training points and targets  $(x_1, y_1) \dots (x_n, y_n)$ , where  $x_i \in \mathbb{R}^n$  and  $y_i \in \{-1, 1\}$ . Consider also a set of testing points T. Intuitively, a testing point  $t_i$  that was wrongly classified must have low confidence and a point that was correctly classified must have high confidence. But what is the lowest value that confidence or credibility can take for a specific training set. This lowest value is referred to as confidence reference (credibility reference) in what follows.

#### 5.1 Geometric analysis of confidence

In this subsection we propose a method to find the confidence reference value knowing only the training set. Consider any two support vectors  $v_+$  and  $v_-$  belonging to classes (+1) and (-1) and having Lagrange multipliers  $\alpha_+$  and  $\alpha_-$ , respectively. Consider a point  $v_t$  which is defined by:

$$v_t = v_- + t(v_+ - v_-),$$

where  $t \in [0, 1]$ . This point is a variable point on the segment enclosed by the two support vectors. When computing the confidence over  $v_t$  the following analysis holds: As  $t$  increases  $v_t$  moves from  $v_-$  to  $v_+$ . To compute the confidence,  $v_t$  is added to the training set first as of class (-1). This will remove  $v_-$  from the support vector set because  $v_t$  is now a support vector of class (-1). Indeed, since  $v_+$  and  $v_-$  are two support vectors from different classes, it is necessary that the hyperplane passes between them. When  $v_t$  is added to the training points as in class (-1), then  $v_-$ ,  $v_t$  and  $v_+$  are collinear in that order. Hence the hyperplane is now passing between  $v_t$  and  $v_+$  because they are of different classes, while  $v_t$  and  $v_-$  are of the same class. Thus  $v_-$  is not a support vector anymore. Hence the Lagrange multiplier  $\alpha_-$  is now equal to zero and  $\alpha_t$  has a value greater than the old value of  $\alpha_-$ , because  $v_t$  is now closer to  $v_+$  than  $v_-$ .  $\alpha_+$  also increases for this same reason. As  $v_t$  gets closer to  $v_+$ ,  $\alpha_t$  and  $\alpha_+$  increase until reaching their limit C defined in Eq. (3). C is always reached because at the limit, when  $v_t$  approaches  $v_+$ , the Hessian matrix will have two identical lines with each line representing a point of different class. This will make the Hessian matrix, at the limit, singular. Thus theoretically  $\alpha_t$

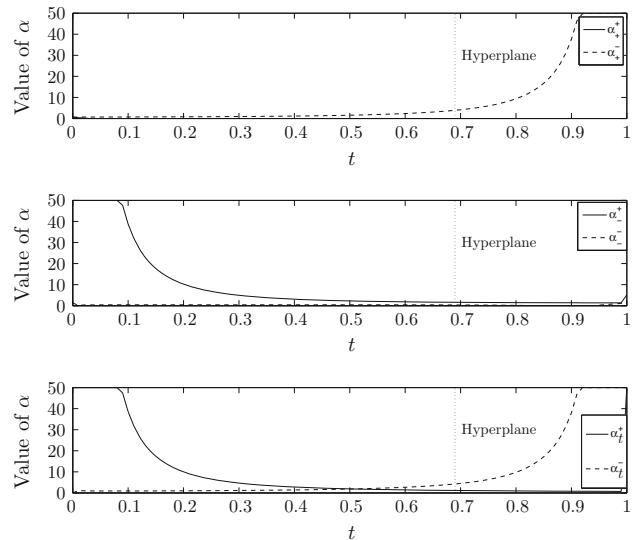
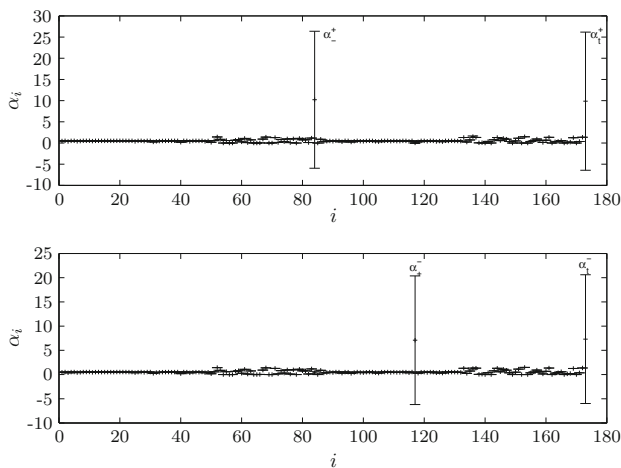


Fig. 1 Variation of  $\alpha_+$ ,  $\alpha_-$  and  $\alpha_t$  as a function of  $t$

and  $\alpha_+$  should be infinite. But since C is their upper limit then  $\alpha_t = \alpha_+ = C$ .

Next  $v_t$  is trained as class (+1). This will remove  $v_+$  from the support vector set for the same reason stated above except that now  $v_t$  is in class (+1). Hence the Lagrange multiplier  $\alpha_+$  is now equal to zero, and  $\alpha_t$  and  $\alpha_-$  have their greatest value C, because  $v_t$  is now very close to  $v_-$ . As  $v_t$  moves away from  $v_-$ ,  $\alpha_t$  and  $\alpha_-$  decrease until  $\alpha_t$  reaches 0 when  $t = 1$  and  $\alpha_-$  has its original value. Figure 1 shows the analysis presented above. As we can see  $\alpha_+^-$ , which is the value of  $\alpha_+$  when  $v_t$  is trained as class (-1), increases from its initial value  $\alpha_{+0}$  to C; while  $\alpha_+^+$ , which is the value of  $\alpha_+$  when  $v_t$  is trained as class (+1), goes to zero.  $\alpha_-^+$ , which is the value of  $\alpha_-$  when  $v_t$  is trained as class (+1), decreases from C to its initial value  $\alpha_{-0}$ ; while  $\alpha_-^-$ , which is the value of  $\alpha_-$  when  $v_t$  is trained as class (-1), goes to zero. As for  $\alpha_t^+$ , which is the value of  $\alpha_t$  when  $v_t$  is trained as (+1), decreases from C to  $\alpha_{+0}$ , while  $\alpha_t^-$ , which is the value of  $\alpha_t$  when  $v_t$  is trained as (-1), increases from  $\alpha_{-0}$  to C.

The objective is to find the lowest confidence while moving from  $v_-$  to  $v_+$ . As the variable point moves from  $v_-$ , its decision is -1; hence the confidence criterion is  $\zeta = d(z)(VC_{-1} - VC_1) = VC_1 - VC_{-1}$ . Minimizing  $\zeta$  minimizes the confidence. Hence to minimize the confidence, the point that maximizes  $VC_{-1}$  and minimizes  $VC_1$ , while  $d(z) = -1$ , must be chosen. Note that by varying  $t$ , only  $\alpha_+$ ,  $\alpha_-$  and  $\alpha_t$  change values and all the other  $\alpha_i$  exhibit a negligible deviation from their original values because  $v_t$  is closest to  $v_+$  and  $v_-$ ; thus their corresponding Lagrange multipliers will have the higher variation. This is also shown experimentally in Fig. 2. The figure represents the mean of  $\alpha_i$  for all support vectors while  $v_t$  is moving from  $v_-$  to  $v_+$ , with the error bars



**Fig. 2** Variability of  $\alpha_i$  while  $v_t$  moves from  $v_-$  to  $v_+$

around each point corresponding to plus and minus standard deviation. As we can see only  $\alpha_+^+$ ,  $\alpha_+^-$ ,  $\alpha_t^+$  and  $\alpha_t^-$  have a considerable standard deviation. Thus to maximize  $VC_{-1}$ , the values  $\alpha_+^-$ ,  $\alpha_+^+$  and  $\alpha_t^-$  must be maximal. Since  $\alpha_+^- = 0$  for any  $t > 0$ , and  $\alpha_+^+$  and  $\alpha_t^-$  are an increasing function of  $t$ , the point that maximizes  $VC_{-1}$  is the point that corresponds to the highest value of  $t$  before changing decision from  $-1$  to  $+1$ . Denote by  $t_h$  the point on the hyperplane; then

$$t_{\max} = \lim_{t \rightarrow t_h^-} t$$

To minimize  $VC_1$  the same procedure is used: the values  $\alpha_+^+$ ,  $\alpha_+^-$  and  $\alpha_t^+$  must be minimal. Since  $\alpha_+^+ = 0$  and  $\alpha_+^-$  and  $\alpha_t^+$  are a decreasing function of  $t$ , the point that minimizes  $VC_1$  corresponds to the highest value of  $t$  before changing decision from  $-1$  to  $1$ , which is the same  $t_{\max}$  defined above. Thus the point lying on the hyperplane between two support vectors of different classes have the lowest confidence. Between all support vectors from different classes, the two support vectors that yield a point on the hyperplane that has the lowest confidence between all other points on the hyperplane, are the two vectors that correspond to the highest increasing rate of  $\alpha_+^-$  and  $\alpha_t^-$  and the highest decreasing rate of  $\alpha_+^+$  and  $\alpha_t^+$ .

This can be seen from Fig. 1: the intersection of  $\alpha_+^-$  with the hyperplane must be high, as for  $\alpha_+^+$  its intersection must be low. The same holds for  $\alpha_t$  where the intersection point of  $\alpha_t^-$  with the hyperplane must be high and the intersection of  $\alpha_t^+$  with the hyperplane must be low. To achieve this the gradient of the curves must be high. To find the reference values of both regions, one can use the exhaustive search method and try all possible combinations of support vectors starting from the negative region toward the positive region and then repeat this procedure starting from positive to negative and find the lowest confidences of both regions. If the number of support vectors of negative type is  $N_n$  and of positive type is  $N_p$ , this

will require  $2N_nN_p$  search paths. By noting the following two propositions this number can be reduced to  $N_nN_p$ :

**Proposition 1** Let  $T_{\max}^-$  be the point in space corresponding to  $t_{\max}$  on the negative side of the hyperplane and  $T_{\max}^+$  be the point corresponding to  $t_{\max}$  on the positive side of the hyperplane; then the confidence between these two points is related by

$$C_f(T_{\max}^-) + C_f(T_{\max}^+) = 1$$

*Proof* The criterion for  $t_{\max}$  on the negative side is given by  $\zeta^- = d(z)(VC_{-1} - VC_1) = VC_1 - VC_{-1}$  because  $d = -1$ . For  $T_{\max}^+$  the criterion is the same but with  $d(z)=1$  thus  $\zeta^+ = VC_{-1} - VC_1 = -\zeta^-$ . Thus

$$\begin{aligned} C_f(T_{\max}^-) + C_f(T_{\max}^+) &= \frac{1}{1 + e^{-\zeta^-}} + \frac{1}{1 + e^{\zeta^-}} \\ &= \frac{1}{1 + e^{-\zeta^-}} + \frac{e^{-\zeta^-}}{1 + e^{-\zeta^-}} \\ &= 1 \end{aligned}$$

□

**Proposition 2** Let  $C_{f,\min}^-$  be the minimal confidence achieved on the hyperplane from the negative side between any 2 support vectors of different classes and  $C_{f,\max}^-$  the maximum confidence on the hyperplane from the negative side between any 2 support vectors; then the reference confidence values for the negative and positive regions are given by

$$\begin{aligned} C_{\text{ref}}^- &= C_{f,\min}^- \\ C_{\text{ref}}^+ &= 1 - C_{f,\max}^- \end{aligned}$$

*Proof* Because of the discontinuity of the confidence on the hyperplane if a hyperplane has a high confidence on one side it will have a low confidence on the other (from proposition 1). For the negative side reference confidence since  $C_{f,\min}^-$  is by definition the minimal confidence achieved on the hyperplane from the negative side between any 2 support vectors of different classes, it is the reference point for the negative side. Since  $C_{f,\max}^-$  is the highest confidence achieved on the hyperplane on the negative side, from proposition 1 it will have the lowest confidence from the other side; hence it is the reference confidence for the positive region. □

Thus the search need only be exhaustive from one side to deduce the reference points for both regions, which reduces the number of paths to  $N_nN_p$ .

### 5.2 Geometric analysis of credibility

The same analysis for confidence can be applied to credibility. Remember that the criterion for credibility is given

by  $\xi = (VC(d) - VC_0)$  and that credibility decreases as  $\xi$  increases. Hence the minimum credibility is achieved when  $VC(d)$  is maximum. If  $d = -1$  then the minimum credibility is achieved when  $VC(-1)$  is maximum which is achieved on the hyperplane from the analysis done for confidence. If  $d = 1$  then the minimum credibility is achieved when  $VC(1)$  is maximum which is also achieved on the hyperplane from the analysis made for confidence. Thus the reference point for credibility is also found on the hyperplane. Note that it is not necessary that the confidence reference point be the same as the credibility reference point because for confidence the maximization was over  $VC_{-1} - VC_1$  while for credibility the maximization is over  $VC_{-1}$  and  $VC_1$  separately. The same procedure to find the reference point can be applied for both regions. Also note that starting from close points will yield the highest rate as well as considering the points having the highest  $\alpha_0$  which is the initial starting point will lead to a faster convergence to the reference point.

### 5.3 Confidence and credibility limit theorem

The interest in this subsection is on points  $\Omega$  that are away from any support vectors. Formally this can be written as

$$\|x_k - \Omega\| \rightarrow \infty \quad \forall k \in \{1, 2, \dots, n\}$$

The following theorem will show that with a good choice of the kernel function, all such points have the same confidence and credibility levels in their regions. Before stating and proving the theorem, the following definition, notation and lemmas are necessary.

**Definition 3** A kernel function is said to have the Radial Basis Function (RBF) properties if it satisfies the following properties:

1.  $0 < K(x, y) \leq 1$ .
2.  $K(x, x) = 1$ .
3.  $\lim_{\|x-y\| \rightarrow \infty} K(x, y) \rightarrow 0$ .

In the remainder of this paper, only the kernel that has RBF properties is considered.

**Notation** The following notations are used in the next lemmas and theorem:

- $H_1(z)$  denotes the Hessian matrix when the testing point  $z$  is appended to the training points as of class +1.
- $H_{-1}(z)$  denotes the Hessian matrix when the testing point  $z$  is appended to the training points as of class -1.

**Lemma 1** 1. If  $\|\Omega - x_k\|^2 \rightarrow \infty, \forall k$  then  $H_{-1}(\Omega) = H_1(\Omega)$  and hence for points lying at infinite distance, the subscript is omitted.

2. If  $\Omega_1$  and  $\Omega_2$  are two unseen vectors such that  $\|\Omega_1 - x_k\|^2 \rightarrow \infty$  and  $\|\Omega_2 - x_k\|^2 \rightarrow \infty, \forall k$  then  $H(\Omega_1) = H(\Omega_2)$ .

*Proof* 1. If  $H$  is the Hessian matrix before adding  $\Omega$  to the training set, then  $H$  is  $(n \times n)$ . While  $H_{-1}(\Omega)$  is obtained by adding one row and column to  $H$ . So  $H_{-1}(\Omega)$  is an  $(n + 1) \times (n + 1)$  matrix identical to  $H$  except for the added row and column. (This follows from the definition of  $h_{ij}$ : if  $i, j < n + 1, h_{ij}$  will not be affected by adding  $\Omega$  to the training set). Since  $H$  is symmetric it is sufficient to compute the last row only:  $h_{(n+1)i} = y_\Omega y_i K(\Omega, x_i) = 0 \quad \forall i < (n + 1)$  because  $\|\Omega - x_i\|^2 \rightarrow \infty, \forall i$  and thus  $K(\Omega, x_i) \rightarrow 0 \quad \forall i$ . The last term is given by  $h_{(n+1)(n+1)} = y_\Omega y_\Omega K(\Omega, \Omega) = 1$ . Thus  $H_{-1}(\Omega)$  is the extension of  $H$  by adding a row of zeros and a column of zeros except for the diagonal term which ends up to be equal to 1. The same reasoning holds for  $H_1(\Omega)$ , the only difference now is that  $y_\Omega = 1$  which makes no difference on the diagonal element because  $y_\Omega$  appears as a square thus  $h_{(n+1)(n+1)} = y_\Omega y_\Omega K(z, z) = 1$ . And hence  $H_{-1}(\Omega) = H_1(\Omega) = H(\Omega)$ . the above explanation can be seen on the following two matrices.

$$\overbrace{\begin{pmatrix} 1 & \alpha_{12} & \dots & \alpha_{1n} \\ \vdots & & & \\ \alpha_{n1} & \alpha_{n2} & \dots & 1 \end{pmatrix}}^{H(\Omega)}, \quad \overbrace{\begin{pmatrix} 1 & \dots & \alpha_{1n} & 0 \\ \vdots & & & \\ \alpha_{n1} & \dots & 1 & 0 \\ 0 & \dots & 0 & 1 \end{pmatrix}}^{H_{-1}(\Omega)=H_1(\Omega)} \quad (21)$$

2. The proof follows from part 1, as it has been shown for any vector  $\Omega$  that satisfies  $\|\Omega - x_i\|^2 \rightarrow \infty \quad \forall i$ , the Hessian matrix that corresponds to adding this vector to the training set is obtained by adding a row and a column of zeros to the original Hessian matrix and only the diagonal element is equal to 1. And thus  $H(\Omega_1) = H(\Omega_2)$ . □

The Hessian matrix remains the same if  $\Omega$  was trained as  $(-1)$  or  $(1)$  and is the same for any 2 points that are far away from the training set. This result is used in the proof of the confidence limit theorem. Another key result is related to the radius of the smallest sphere containing all training set which is computed in the following lemma:

**Lemma 2** Let  $\Omega_1$  and  $\Omega_2$  be two vectors such that  $\|\Omega_1 - x_i\|^2 \rightarrow \infty$  and  $\|\Omega_2 - x_i\|^2 \rightarrow \infty, \forall i$  then:

$$D_{\Omega_1}^2 = D_{\Omega_2}^2 = D_{max} = \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j K(x_i, x_j) + 1, \quad (22)$$

where the  $\lambda_i$  is the solution of Eq. (15).

*Proof* To find the sphere with the smallest radius  $D$ , we continue the analysis of Eq. (16) by considering the RBF properties of the kernel function; hence

$$D^2 = \max_{x_k} \left\{ \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j K(x_i, x_j) - 2 \sum_{i=1}^n \lambda_i K(x_i, x_k) + \underbrace{K(x_k, x_k)}_{=1} \right\}$$

Only one term depends on  $x_k$ . Thus to maximize  $D$ , the term that depends on  $x_k$  has to be minimized, and since  $K(x, y) > 0$  and  $\lambda_i > 0$  then the maximum is achieved when  $K(x_i, x_k) = 0$ . To compute  $D_{\Omega_1}$  it is obvious that the radius will be obtained by replacing  $x_k$  by  $\Omega_1$  because  $\|x_i - \Omega_1\|^2 \rightarrow \infty$  and thus  $K(x_i, \Omega_1) \rightarrow 0$  and hence  $D^2 = D_{max}$ . The same reasoning applies when  $\Omega_2$  is added to the training points instead of  $\Omega_1$  and the equation of the radius will not change and hence

$$D_{\Omega_1}^2 = D_{\Omega_2}^2 = \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j K(x_i, x_j) + 1.$$

We now state and prove the confidence limit theorem.

**Theorem 1** (Confidence Limit Theorem) *Let  $\Omega_1$  and  $\Omega_2$  be two unseen vectors such that  $\|\Omega_1 - x_k\|^2 \rightarrow \infty$  and  $\|\Omega_2 - x_k\|^2 \rightarrow \infty, \forall k \in \{1, 2, \dots, n\}$  and a kernel function having RBF properties; then*

$$|\zeta_{\Omega_1}| = |\zeta_{\Omega_2}| = |\zeta_L|,$$

where  $\zeta_L$  is a finite limit.

*Proof* Consider  $\Omega_1$  and  $\Omega_2$  such that  $\|\Omega_1 - x_i\|^2 \rightarrow \infty$  and  $\|\Omega_2 - x_i\|^2 \rightarrow \infty, \forall i$ , if  $\Omega_i$  is appended to the training set then:

$$|\zeta_{\Omega_i}| = |VC_{-1} - VC_1| = D_{\Omega_i}^2 \left( \|w(\Omega_i)\|_{-1}^2 - \|w(\Omega_i)\|_1^2 \right),$$

where  $\|w(\Omega_i)\|_{-1}^2$  is the norm of the weight of the hyperplane when  $\Omega_i$  is considered to be of class  $(-1)$ , and  $\|w(\Omega_i)\|_1^2$  is the norm of the weight of the hyperplane when  $\Omega_i$  is considered to be of class  $(1)$ . Lemma 2 shows that  $D_{\Omega_1}^2 = D_{\Omega_2}^2 = D_{max}$  so in what follows we will concentrate only on  $\|w\|^2$ .

To find  $\|w(\Omega_i)\|_{-1}^2$  and  $\|w(\Omega_i)\|_1^2$  we need to solve the minimization problem in both cases [When  $\Omega_i$  is appended to the training set and considered as  $(-1)$  and when it is considered as  $(+1)$ ].

If  $\Omega_1$  is considered to be of class  $(-1)$ , then the minimization problem is given by

$$L_d(\alpha) = 0.5\alpha^T H_{\Omega_1} \alpha - f^T \alpha \tag{23}$$

subject to the following constraints:

$$y_{-1}^T \alpha = 0, \\ 0 \leq \alpha \leq C.$$

Solving the above minimization will yield  $\|w(\Omega_1)\|_{-1}^2$ . If  $\Omega_1$  is considered to be of class  $(+1)$ , then the minimization problem is given by

$$L_d(\alpha) = 0.5\alpha^T H_{\Omega_1} \alpha - f^T \alpha \tag{24}$$

subject to the following constraints:

$$y_1^T \alpha = 0, \\ 0 \leq \alpha \leq C.$$

Solving the above minimization will yield  $\|w(\Omega_1)\|_1^2$ . If  $\Omega_2$  is considered to be of class  $(-1)$ , then the minimization problem is given by

$$L_d(\alpha) = 0.5\alpha^T H_{\Omega_2} \alpha - f^T \alpha \tag{25}$$

subject to the following constraints:

$$y_{-1}^T \alpha = 0, \\ 0 \leq \alpha \leq C.$$

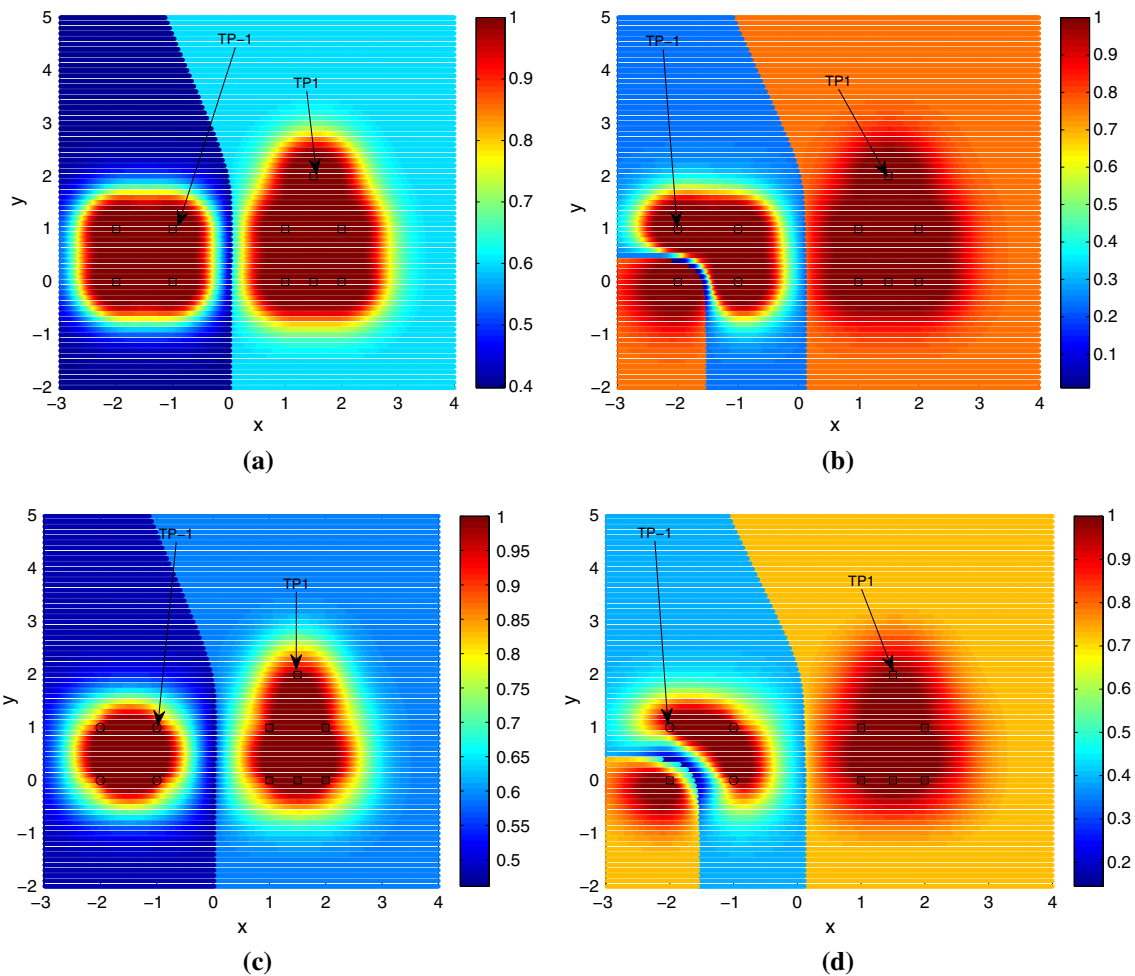
Solving the above minimization will yield  $\|w(\Omega_2)\|_{-1}^2$ . If  $\Omega_2$  is considered to be of class  $(+1)$ , then the minimization problem is given by

$$L_d(\alpha) = 0.5\alpha^T H_{\Omega_2} \alpha - f^T \alpha \tag{26}$$

subject to the following constraints:

$$y_1^T \alpha = 0, \\ 0 \leq \alpha \leq C.$$

Solving the above minimization will yield  $\|w(\Omega_2)\|_1^2$ . Lemma 1 shows that  $H_{\Omega_1} = H_{\Omega_2}$ ; thus Eqs. (23) and (25) are identical and thus  $\|w(\Omega_1)\|_{-1}^2 = \|w(\Omega_2)\|_{-1}^2$ . The same reasoning shows that Eqs. (24) and (26) are also identical and hence  $\|w(\Omega_1)\|_1^2 = \|w(\Omega_2)\|_1^2$ . This result combined with the fact that  $D_{\Omega_1}^2 = D_{\Omega_2}^2 = D_{max}$  implies the result of the theorem.  $\zeta_L$  is finite simply because  $D_{\Omega_1}^2 = D_{\Omega_2}^2 = D_{max}$  is finite and also because the number of training points is finite and  $\alpha_i \leq C$  which implies by (13) that  $\|w\|^2$  is finite and hence  $\zeta_L$  is finite.



**Fig. 3** Confidence and credibility variation. **a** Linearly separable dataset confidence map, **b** non-linearly separable dataset confidence map, **c** linearly separable dataset credibility map, **d** non-linearly separable dataset credibility map

This proof shows that for far away points their VC is constant and hence all far away points have the same credibility in their region, so class  $-1$  has its own limit confidence and limit credibility; while class  $+1$  has another limit of confidence and credibility. Note that since it is necessary to have a hyperplane separating the far away points of different regions, it is possible to state that the sum of the confidence limit of both sides is equal to 1 from Proposition 1. Figure 3 shows the variation of the confidence and credibility levels over a two-dimensional dataset. It shows clearly the limit of the credibility and confidence over the far points as well as the reference points found on the hyperplane. It is also clear in the non-linear data that the confidence and credibility on the hyperplane that separate the closer points are less than their values on the hyperplane that separates the points that are further apart. In this example a Gaussian kernel function is used for the linearly separable dataset and for the non-linearly separable. Although for the linearly separable data, the linear kernel function seems more suitable, we used the Gaussian because it is visually more intuitive. One limitation of this method

is the relatively high computational complexity. To compute the confidence for one sample, SVM has to be trained twice. The computational complexity of the training phase of SVM is  $\mathcal{O}(n^3)$ ; hence to calculate the confidence and credibility levels we need  $2\mathcal{O}(n^3)$ . Figure 3 suggests a certain correlation between confidence and credibility; hence the following theoretical study on the correlation between confidence and credibility was performed.

#### 5.4 Correlation between confidence and credibility

We now present a theoretical study of the correlation between confidence and credibility in which it is shown how the correlation varies with the training points. The study is done on the criterion of the confidence and credibility. Since the criterion of confidence is:  $sign(d)(VC_{-1} - VC_1)$  this could be written as  $\overline{VC} - VC$ , where VC represents the VC of the decision and  $\overline{VC}$  represents the VC of not the decision. The criterion of the credibility is  $VC(d) - VC_0$ . It could be written as  $VC - VC_0$ . Let  $X = \overline{VC} - VC$  and  $Y = VC - VC_0$ .

Then  $E[X] = \bar{\mu} - \mu$  and  $E[Y] = \mu - VC_0$  because  $VC_0$  is a constant. The correlation coefficient is given by

$$r_{XY}^2 = \frac{\text{cov}(X, Y)^2}{\text{Var}(X)\text{Var}(Y)}$$

$$\begin{aligned} \text{cov}(X, Y) &= E[(\overline{VC} - VC - \bar{\mu} - \mu)(VC - \mu)] \\ &= -\mu\bar{\mu} + \mu^2 + \mu\bar{\mu} - \mu^2 + E[\overline{VC}VC] - E[VC^2] \\ &\quad - \bar{\mu}\mu + \mu^2 \\ &= E[\overline{VC} - VC] - \bar{\mu}\mu - (E[VC^2] - \mu^2) \\ &= \text{cov}(\overline{VC}, VC) - \text{Var}(VC) \end{aligned}$$

and

$$\begin{aligned} \text{Var}(X) &= E[X^2] - E[X]^2 \\ &= E[\overline{VC}^2 + VC^2 - 2VC\overline{VC}] - \bar{\mu}^2 - \mu^2 + 2\bar{\mu}\mu \\ &= \text{Var}(\overline{VC}) + \text{Var}(VC) - 2\text{cov}(VC, \overline{VC}) \end{aligned}$$

and  $\text{Var}(Y) = \text{Var}(VC)$  Thus

$$r_{xy}^2 = \frac{\text{cov}(VC, \overline{VC}) + \text{Var}(VC^2) - 2\text{Var}(VC)\text{cov}(VC, \overline{VC})}{\text{Var}(\overline{VC})\text{Var}(VC) + \text{Var}(VC^2) - 2\text{Var}(VC)\text{cov}(VC, \overline{VC})}$$

Then we can deduce that to have  $r_{xy}^2 = 1$  we must have

$$\text{Cov}(VC, \overline{VC})^2 = \text{Var}(VC)\text{Var}(\overline{VC})$$

which means

$$|r_{VC, \overline{VC}}| = 1,$$

which implies that  $VC$  and  $\overline{VC}$  are linearly related. This occurs when the data used are linearly separable or are very simple to classify which implies that whenever  $VC$  increases,  $\overline{VC}$  decreases thus they are negatively correlated. But if the data used have high randomness, then probably  $VC$  and  $\overline{VC}$  are not linearly correlated hence the importance of the 2 developed measures in difficult datasets. Note that the confidence and credibility measures are of lesser importance in easy datasets; hence the correlation poses limited drawback to our measures.

Having defined the expression of the confidence and credibility measures, they are applied on the well-known dataset of the US postal digit recognition database (Cun et al. 1990). The confidence and credibility measures are also applied for agitation detection and for the Wisconsin breast cancer dataset. The next section describes the setup and results.

## 6 Experiments and results

This section presents the results of the proposed confidence measure and credibility on different datasets. The results are

compared to those provided by Vovk's method and by a modified version of Vovk's method they are also compared to Platt's method which is a well-known benchmark for confidence measure in SVM. The classification in Vovk's method is independent of SVM's classification which is a drawback because SVM has always a higher accuracy as stated in his paper and as the following results show. We modified Vovk's method by taking the classification of SVM as the final decision for his method and then computed the confidence and credibility according to his definition. This modification actually improved Vovk's confidence and credibility results. However, as will be shown our proposed method outperforms both Vovk's original method and Vovk's improved method. Good confidence and credibility measures must have high values for correctly classified points and low value for the points that were incorrectly classified. One possible method to measure their performance is the Brier score (1950) which is defined as:

$$BS = \frac{1}{N} \sum_{i=1}^N (f_i - o_i)^2,$$

where  $f_i$  is the outputted measure and  $o_i$  is a binary variable that takes the value 1 if the event is true and 0 if it is false. When comparing two measures, the one having the lower BS value has higher correlation with the event. However, this method works only on measures representing a probability distribution which is neither the case of our measures nor of Vovk's measures. So to measure the performance of a measure, we will find its average over correctly classified samples and over wrongly classified ones. A good measure should have a larger difference between these two averages as well as a small standard deviation.

### 6.1 US postal data

The proposed measures are tested using the US postal digit recognition problem. This is a well-known dataset that was used by Cun et al. (1990). The addressed problem is the binary pattern recognition problem. Many papers dealt with the recognition problem of the two digits "2" and "7" together and "3" and "8" together and hence the choice of these 2 sets of two digits to apply our proposed measures. The numbers are represented using  $16 \times 16$  pixels that represent the normalized gray scale of each pixel. Each number has multiple samples coming from different people's handwriting. In total 1,720 samples for both numbers "2" and "7" are studied. Similarly 1,720 samples for the numbers "3" and "8" are studied with another classifier. The  $K$  fold cross-validation technique is used to evaluate the performance of the proposed confidence measure. The 1,720 samples are divided into 10 groups of 172 samples each. The tenfold cross-validation is carried out by taking onefold as the training set and the

remaining ninefold as testing set. One kernel function that satisfies the RBF property is the Gaussian kernel function defined as

$$K(x_i, x_j) = \exp\left(-\frac{1}{2\sigma^2}(x_i - x_j)^T(x_i - x_j)\right)$$

To choose the best parameter  $1/\sigma^2$  for the SVM, a subset of 2 groups is considered. A onefold cross-validation was conducted where  $1/\sigma^2$  was varied over a large range. The value of  $1/\sigma^2$  that gave the best average performance over the 2 groups was chosen in the larger tenfold cross-validation procedure. The value found is  $\sigma = \sqrt{5}$  for the 2–7 classifier and  $\sigma = \sqrt{2.5}$  for the 3–8 classifier. It is worth noting that no training point was used at the same time for training and testing; this is a standard procedure to avoid biasing the accuracy of the testing phase. As stated above a good way to evaluate the performance of the confidence and credibility measures is by their BS; we also evaluate their average over the points that were classified correctly and compare it with their average over the points that were misclassified by SVM. The average has to be taken over all ten permutations. This method is compared to Vovk’s method described in (1999) and to its modified version described earlier. It is also compared with Platt’s log sigmoid confidence. The log sigmoid parameters A and B were found by using the technique described in his paper and using the algorithm that he provided (Platt 1999).

The last column in Table 1 shows that the difference between the average confidence over correctly classified points and the incorrectly classified ones by our proposed confidence measure (0.52) is higher than the one introduced by Vovk (0.01) and by its modified version (0.17) as well as the confidence introduced by Platt (0.17). The same holds for our proposed credibility measure where the difference obtained by our method is 0.35 in comparison with the one given by Vovk (0.28) and by its modified version (0.25).

The same reasoning in Table 2 shows that the difference between the average confidence over correctly classified points and the incorrectly classified ones (last column) by our

**Table 1** Confidence and credibility for US Postal 2–7

Method	Mean <i>T</i>	Std <i>T</i>	Mean <i>F</i>	Std <i>F</i>	Mean( <i>T</i> ) –Mean( <i>F</i> )
VC credibility	0.763	0.159	0.419	0.148	<b>0.35</b>
VC confidence	0.825	0.195	0.306	0.242	<b>0.52</b>
Vovk credibility	0.524	0.269	0.242	0.062	0.28
Vovk confidence	0.991	0.023	0.983	0.022	0.01
Vovk SVM credibility	0.511	0.287	0.266	0.118	0.25
Vovk SVM confidence	0.978	0.057	0.807	0.022	0.17
Platt’s confidence	0.39	0.14	0.22	0.09	0.17

Bold values are the best ones achieved

**Table 2** Confidence and credibility for US Postal 3–8

Method	Mean <i>T</i>	Std <i>T</i>	Mean <i>F</i>	Std <i>F</i>	Mean( <i>T</i> ) –Mean( <i>F</i> )
VC credibility	0.72	0.11	0.409	0.09	<b>0.32</b>
VC confidence	0.75	0.14	0.5	0.1	<b>0.25</b>
Vovk credibility	0.5	0.154	0.262	0.07	0.24
Vovk confidence	0.54	0.015	0.54	0.017	0
Vovk SVM credibility	0.48	0.24	0.24	0.13	0.24
Vovk SVM confidence	0.56	0.078	0.402	0.082	0.16
Platt’s confidence	0.32	0.17	0.1	0.05	0.22

Bold values are the best ones achieved

proposed confidence measure (0.25) is higher than the one introduced by Vovk (0) and by its modified version (0.16) as well as the confidence introduced by Platt (0.22). The same holds for our proposed credibility measure where the difference obtained by our method is 0.32 in comparison with the one given by Vovk (0.24) and by its modified version (0.24).

### 6.2 Agitation detection

The problem of agitation detection is a well-studied area specially for people with dementia. Subject independent agitation detection is based on training the support vector machines algorithm over a very limited set of subjects and then test it on the remaining subjects to detect agitation independently of the tested subject. This approach has been studied by our team in Sakr et al. (2008, 2009), where normal SVM and distance based multi-level SVM were used respectively and in Sakr et al. (2010), where the binary confidence measure was used. In this experiment we intend to show that correctly classified points by SVM have on average higher confidence than the wrongly classified points. The features used are the skin temperature, the galvanic skin response and the heart rate variability (HRV). From the heart rate, the inter-beat interval (IBI) was extracted. To measure and record the physical features, the following sensors were used: Polar exercise heart rate monitor from Vernier, a 1,000-ohms platinum (RTD) from Omega and electrodes that wrap around the fingers for monitoring galvanic skin response. The RTD sensor changes its resistance with the skin temperature of the subject. The change in resistance is converted into temperature change using the Callendar-Van Dusen equation:

$$R_t = R_0 + R_0\alpha \left[ t - \delta \left( \frac{t}{100} - 1 \right) \left( \frac{t}{100} \right) - \beta \left( \frac{t}{100} - 1 \right) \left( \frac{t}{100} \right)^3 \right] \tag{27}$$

Since the measured temperature is always above 0, only the Callendar coefficient  $\delta$  is used, while the Van Dusen coefficient  $\beta = 0$  for positive temperatures.

The experimental procedure is as follows: the subject places the sensors around his body. He then undergoes the trait scale state-trait anxiety inventory (T-STAI) (Spielberger et al. 1970). The trait anxiety scale is one of the two subscales of the full-form STAI developed by Spielberger to measure anxiety in adults. It is one of the most frequently used measures of anxiety in applied psychology research and has been shown to be a reliable and sensitive measure of anxiety. Subjects were asked to fill the T-STAI before and after the Stroop test. When undergoing the Stroop test, all the signals are being recorded on the same machine where the Stroop test is running (Stroop 1935). A sample is generated per second during the 4-min test, which yields a total number of 240 samples per subject. For details about the data capture protocol and the Stroop test, consult our previous work (Sakr et al. 2010).

In total 58 subjects were tested. The proposed confidence and credibility measures are tested using the  $K$  fold cross-validation technique. The 58 subjects are subdivided in groups of 2 to form 29-fold. In general the cross-validation is carried out by taking 28-fold as the training set and the remaining fold as the validation set. The final result is the average of all  $K$  permutations. To illustrate the robustness of our measures, only onefold is taken as the training set and the remaining 28-fold are taken as the testing set. The same kernel function used for the US postal data is used for agitation detection.

To choose the best parameter  $1/\sigma^2$  for the SVM, a subset of 12 subjects is considered. A sixfold cross-validation was conducted where  $1/\sigma^2$  was varied over a large range. The value of  $1/\sigma^2$  that gave the best average performance over the 12 subjects was chosen in the larger 29-fold cross-validation procedure. The value for agitation detection is  $\sigma = 0.25$ . It is worth noting also that no training point was used at the same time for training and testing. As stated above a good way to evaluate the performance of the confidence and credibility measures is to evaluate their average over the points that were classified correctly and compare it with their average over the points that were misclassified by SVM. The average has to be taken over all 29 permutations. This method is compared to Vovk's method described in Vovk et al. (1999), to its modified version described earlier and to Platt's confidence.

The last column in Table 3 shows that the difference between the average confidence and credibility measures over the correctly classified points and the incorrectly classified points by the proposed measures is higher than the one introduced by Vovk as well as by its modified version; however, the modified version outperforms slightly its unmodified version.

### 6.3 Wisconsin breast cancer dataset

This data set has 569 samples and uses 10 features to predict if a tumor is malignant or benign. The features are computed

**Table 3** Confidence and credibility for agitation detection

Method	Mean $T$	Std $T$	Mean $F$	Std $F$	Mean( $T$ ) −Mean( $F$ )
VC credibility	0.908	0.199	0.781	0.303	<b>0.13</b>
VC confidence	0.934	0.181	0.777	0.337	<b>0.16</b>
Vovk credibility	0.417	0.313	0.386	0.321	0.03
Vovk confidence	0.923	0.0315	0.913	0.0513	0.01
Vovk SVM credibility	0.414	0.314	0.375	0.322	0.04
Vovk SVM confidence	0.920	0.0436	0.901	0.0513	0.02
Platt's confidence	0.39	0.24	0.32	0.28	0.07

Bold values are the best ones achieved

**Table 4** Confidence and credibility for breast cancer dataset

Method	Mean $T$	Std $T$	Mean $F$	Std $F$	Mean( $T$ ) −Mean( $F$ )
VC credibility	0.9	0.14	0.58	0.16	<b>0.32</b>
VC confidence	0.94	0.21	0.777	0.297	<b>0.17</b>
Vovk credibility	0.48	0.22	0.3	0.181	0.18
Vovk confidence	0.95	0.0115	0.94	0.0113	0.01
Vovk SVM credibility	0.47	0.24	0.21	0.19	0.26
Vovk SVM confidence	0.940	0.0836	0.881	0.0713	0.06
Platt's confidence	0.49	0.26	0.41	0.27	0.08

Bold values are the best ones achieved

from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. More information on this dataset and a full description of the features can be found in Street et al. (1992). The same cross-validation procedure was also followed on this dataset and the average confidence and credibility for all the methods are presented. We also performed the same cross-validation procedure to determine the best hyperparameter which in this case was found to be  $\sigma = 45$ .

The last column in Table 4 shows that the difference between the average confidence and credibility measures over the correctly classified points and the incorrectly classified points by the proposed measures is higher than the one introduced by Vovk as well as by its modified version and by Platt's confidence; however, the modified version for Vovk outperforms slightly its unmodified version.

## 7 Conclusion

This paper presented new confidence and credibility measures for support vector machines based on the approximation of the VC bound. This method showed a very good correlation between correctly classified points and high confidence/credibility; it also showed a greater spread of the confidence/credibility measure between correctly classified points

and misclassified points which gives an increased ability to discriminate between classification. The proposed measure outperformed Vovk's method as well as a modified version of his method. Our proposed confidence measure has also outperformed Platt's benchmark log sigmoid confidence measure. However, our method has some limitations that should be addressed in the future. Our method's computational complexity is in the order of  $2\mathcal{O}(n^3)$  which is relatively high especially when the number of training points gets very large. A second limitation is that our defined measures work only with RBF kernel functions. Although RBF kernels are very popular and are widely used in many applications, future studies should focus on extending the theory to cover other kernel functions such as polynomial kernels.

**Acknowledgments** This research was funded by the American University of Beirut University Research Board, Dar Al-Handassah (Shair & Partners) Research Fund and the Rathman (Kadifa) Fund. We would like to thank Dr. Cheryl Riley-Doucet and Dr. Debatosh Debnath from Oakland University for providing the data used in this research.

## References

- Aronszajn N (1950) Introduction to the theory of Hilbert spaces. Reasearch [sic] Foundation, Stillwater
- Brier G (1950) Verification of forecasts expressed in terms of probability. *Mon Weather Rev* 78(1):1–3
- Burges C (1998) A tutorial on support vector machines for pattern recognition. *Data Mining Knowl Discov* 2(2):121–167
- Cun Y, Boser B, Denker J, Howard R, Habbard W, Jackel L, Henderson D (1990) Handwritten digit recognition with a back-propagation network. In: *Advances in neural information processing systems 2*, Morgan Kaufmann Publishers Inc. pp 396–404
- Dawid A (1983) Inference, statistical: I. *Encycl Stat Sci* 4:89–105
- Dumais S, Platt J, Heckerman D, Sahami M (1998) Inductive learning algorithms and representations for text categorization. In: *Proceedings of the seventh international conference on Information and knowledge management*, ACM. pp 148–155
- Ercan, Secil, Kayakutlu G (2014) Patent value analysis using support vector machines. In: *Soft computing*, vol 18. Springer, pp 313–328
- Gamerman A, Vapnik V., Vovk V (1998) Learning by transduction. In: *Uncertainty in Artificial Intelligence*, Morgan Kaufmann. pp 148–155
- Joachims T (1998) Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*. pp 137–142
- Kecman V (2001) *Learning and soft computing: support vector machines, neural networks, and fuzzy logic models*. The MIT press, London
- Kolmogorov A (1968) Three approaches to the quantitative definition of information. *Int J Comput Math* 2(1):157–168
- Li L, Chen J (2006) Emotion recognition using physiological signals. *Advances in Artificial Reality and Tele-Existence*. pp 437–446
- Liu, Zhi and Xu, Shuqiong and Zhang, Yun and Chen, Xin and Chen, Philip CL (2013) Interval type-2 fuzzy kernel based support vector machine algorithm for scene classification of humanoid robot. In: *Soft Computing*, Springer, pp 1–18
- Melluish T, Saunders C, Nouretdinov I, Vovk V (2001) The typicalness framework: a comparison with the bayesian approach. University of London, Royal Holloway
- Mitra P, Murthy C, Pal S (2004) A probabilistic active support vector learning algorithm. *IEEE Trans Pattern Anal Mach Intell* 26(3):413–418
- Platt J (1999) Probabilities for SV machines. *Advances in Neural Information Processing Systems*. pp 61–74
- Sakr G, Elhajj I, Huijjer H (2010) Support Vector Machines to Define and Detect Agitation Transition. *IEEE Transactions on Affective Computing*
- Sakr G, Elhajj I, Huijjer H, Riley-Doucet C, Debnath D (2008) Subject independent agitation detection. In: *IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, AIM 2008. pp 200–204
- Sakr G, Elhajj I, Wejinya U (2009) multi level SVM for subject independent agitation detection. In: *IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, AIM 2009, pp 538–543
- Schölkopf B, Smola A (2002) *Learning with kernels*. The MIT Press
- Spielberger CD, Gorsuch RL, Edward LR (1970) STAI manual for the State-Trait Anxiety Inventory (“self-evaluation questionnaire”). In: *Consulting Psychologists Press*
- Street N, Wolberg W, Mangasarian O (1992) Nuclear feature extraction for breast tumor diagnosis. University of Wisconsin-Madison, Computer Sciences Department
- Stroop JR (1935) Studies of interference in serial verbal reactions. *J Exp Psychol* 18:643–662
- Vapnik V (1995) *The nature of statistical learnin*. Springer, New York
- Vapnik V (1998) *Statistical learning theory*. Wiley, New York
- Vovk V, Gammerman A, Saunders C (1999) Machine-learning applications of algorithmic randomness. In: *Proceedings of the Sixteenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc. pp 444–453