Journal of Computer Science 5 (12): 1003-1008, 2009
ISSN 1549-3636

A New Smooth Support Vector Machine and Its Applications in Diabetes Disease Diagnosis

[1,2]Santi Wulan Purnami, [1]Abdullah Embong, [1]Jasni Mohd Zain and [1]S.P. Rahayu
[1]Faculty of Computer System and Software Engineering, University Malaysia Pahang,
Lebuh Raya Tun Abdul Razak 26300, Kuantan Pahang, Malaysia
[2]Department of Statistics, Institute of Technology Sepuluh Nopember Surabaya
Keputih, Sukolilo, Surabaya, Indonesia 60111

Abstract: Problem statement: Research on Smooth Support Vector Machine (SSVM) is an active field in data mining. Many researchers developed the method to improve accuracy of the result. This study proposed a new SSVM for classification problems. It is called Multiple Knot Spline SSVM (MKS-SSVM). To evaluate the effectiveness of our method, we carried out an experiment on Pima Indian diabetes dataset. The accuracy of previous results of this data still under 80% so far. **Approach:** First, theoretical of MKS-SSVM was presented. Then, application of MKS-SSVM and comparison with SSVM in diabetes disease diagnosis were given. **Results:** Compared to the SSVM, the proposed MKS-SSVM showed better performance in classifying diabetes disease diagnosis with accuracy 93.2%. **Conclusion:** The results of this study showed that the MKS-SSVM was effective to detect diabetes disease diagnosis and this is very promising compared to the previously reported results.

Key words: Smooth support vector machine, diabetes disease diagnosis, classification

INTRODUCTION

Support Vector Machines (SVM) is a new algorithm of data mining technique, recently received increasing popularity in machine learning and statistics community. SVM have been introduced by Vapnik[1] for solving pattern recognition and nonlinear function estimation problems. SVM have become the tool of choice for fundamental classification problem of machine learning and data mining. Unlike traditional methods which minimize the empirical training error, SVM aims at minimizing an upper bound of the generalization error through maximizing the margin between the separating hyperplane and the data. This can be regarded as an approximate implementation of the structure risk minimization principle[1,2].

Although many variants of SVM have been proposed, it is still an active research issue in order to improve for more effective classification. SSVM is a development of SVM that uses smoothing technique. This method was first introduced by Lee[3] in 2001. The basic idea of SSVM is to convert SVM primal formulation to a non smooth unconstrained minimization problem. Since the objective function of this unconstrained optimization problem is not twice differentiable, smoothing function can be applied to smooth this unconstrained problem. Lee[3] have proposed the integral of sigmoid function to approximate the plus function. Then, Yuan have proposed polynomial function[4] and spline function[5].

In this study, we propose a new smooth function to approximate the plus function. This function is called Multiple Knot Spline function which is a modification of the spline function[5]. Then, we applied a new SSVM based on multiple knot spline function to diagnose diabetes disease.

The used data source is Pima Indian diabetes disease taken from the UCI machine learning repository[6]. This dataset is commonly used among researchers that use machine learning methods for diabetes disease classification. The results were also compared with the results of the previous studies reported[8-11].

MATERIALS AND METHODS

In this study, we have used SSVM and MKS-SSVM as material and methods. These are explained as follows:

SSVM: In this session, we describe the outline of reformulation standard SVM[1, 2] to smooth SVM[3]. We

Corresponding Author: Santi Wulan Purnami, Faculty of Computer System and Software Engineering,
University Malaysia Pahang, Lebuh Raya Tun Abdul Razak 26300, Kuantan Pahang, Malaysia

begin with the linear case which can be converted to an unconstrained optimization problem. We consider the problem of classifying m points in the n-dimensional real space R^n, represented by the m×n matrix A, according to membership of each point A_i in the classes 1 or -1 as specified by a given m× m diagonal matrix D with ones or minus ones along its diagonal. For this problem the standard SVM is given by the following quadratic program:

$$\min_{(w,\gamma,y)\in R^{n+1+m}} ve'y + \frac{1}{2}w'w$$
$$\min_{(w,\gamma,y)\in R^{n+1+m}} \quad s.t. D(Aw - e\gamma) + y \geq e \qquad (1)$$
$$y \geq 0$$

Where:
v = A positive weight
y = Slack variable
e = Column vector of one of arbitrary dimension
w = The normal to the bounding planes

$$x'w - \gamma = +1$$
$$x'w - \gamma = -1 \qquad (2)$$

γ determines their location relative to the origin. The linear separating surface is the plane:

$$x'w = \gamma \qquad (3)$$

If the classes are linearly inseparable, the bounding plane as follows:

$$x'w - \gamma + y_i \geq +1, \text{ for } x' = A_i \text{ and } D_{ii} = +1,$$
$$x'w - \gamma - y_i \leq -1, \text{ for } x' = A_i \text{ and } D_{ii} = -1 \qquad (4)$$

These constraints (4) can be written as a single matrix equation as follows:

$$D(Aw - e\gamma) + y \geq e \qquad (5)$$

In the SSVM approach[3], the modified SVM problem is yielded as follows:

$$\min_{(w,\gamma,y)\in R^{n+1+m}} \frac{v}{2}y'y + \frac{1}{2}(w'w + \gamma^2)$$
$$s.t. D(Aw - e\gamma) + y \geq e \qquad (6)$$
$$y \geq e$$

The constraint in Eq. 6, can be written by:

$$y = (e - D(Aw - e\gamma))_+ \qquad (7)$$

Thus, we can replace y in constraint (6) by (7) and convert the SVM problem (6) into an equivalent SVM which is an unconstrained optimization problem as follows:

$$\min_{(w,\gamma)} \frac{v}{2}\left\|(e - D(Aw - e\gamma))_+\right\|_2^2 + \frac{1}{2}(w'w + \gamma^2) \qquad (8)$$

The plus function $(x)_+$, is defined as:

$$(x)_+ = \max\{0, x_i\}, \; i = 1,2,3\ldots n \qquad (9)$$

The objective function in (8) is undifferentiable and unsmooth. Therefore, it cannot be solved using conventional optimization method, because it always requires that the objective function's gradient and Hessian matrix.

Lee *et al.*[3] applies the smoothing techniques and replace x_+ by the integral of the sigmoid function:

$$p(x,\alpha) = x + \frac{1}{\alpha}\log(1 + \varepsilon^{-\alpha x}), \; \alpha > 0 \qquad (10)$$

This p function with a smoothing parameter α is used here to replace the plus function of (8) to obtain a Smooth Support Vector Machine (SSVM):

$$\min_{(w,\gamma)\in R^{n+1}} \frac{v}{2}\left\|p(e - D(Aw - e\gamma), \alpha)\right\|_2^2 + \frac{1}{2}(w'w + \gamma^2) \qquad (11)$$

The solution of problem (6) is obtained by solving problem (11) with α approaching infinity. The problem (11) can be solved using a Newton-Armijo algorithm[3].

For nonlinear un-separable problem requires choosing kernel function K to reflect the input space into another space. This model was derived from generalized support vector machines[7]. So the problem (6) can be approximated as following:

$$\min_{(u,\gamma,y)} \frac{v}{2}y'y + \frac{1}{2}(u'u + \gamma^2)$$
$$s.t. D(K(A,A')Du - e\gamma) + y \geq e \qquad (12)$$
$$y \geq 0$$

Same as previous, it is obtained the SSVM for inseparable problem:

$$\min_{(u,\gamma)} \frac{v}{2}\left\|p(e - D(K(A,A')Du - e\gamma), \alpha)\right\|_2^2 + \frac{1}{2}(u'u + \gamma^2) \qquad (13)$$

where, $K(A,A')$ is a kernel map from $R^{mxn} x R^{nxm}$ to R^{mxm}. We can also apply the Newton-Armijo Algorithm directly to solve (13).

Multiple Knot Spline-SSVM (MKS-SSVM): Smooth Support Vector Machines (SSVM) which had been proposed by Lee *et al.*[3] is very important and significant result to SVM because many algorithms can be used to solve it. In SSVM, the smooth function in objective function (13) is the integral of sigmoid function (9). In this study, we propose a new smooth function which called Multiple Knot Spline (MKS) function. The formulation and performance analysis of new smooth function and how to construct to new SSVM will be described as follows:

Multiple knot spline function: In this study, a new smooth function was proposed. It is a Multiple Knot Spline function as following:

$$m(x)=\begin{cases}0, & x<\dfrac{-1}{k}\\ \dfrac{k^2x^3}{6}+\dfrac{kx^2}{2}+\dfrac{1}{2}x+\dfrac{1}{6k}, & \dfrac{-1}{k}\le x<\dfrac{-2}{5k}\\ \dfrac{25}{18}k^2x^3+\dfrac{5}{4}kx^2+\dfrac{4}{9}x+\dfrac{1}{10k}, & \dfrac{-2}{k}\le x<\dfrac{2}{5k}\\ -\dfrac{k^2x^3}{6}+\dfrac{kx^2}{2}+\dfrac{1}{2}x+\dfrac{1}{6k}, & \dfrac{2}{5k}\le x<\dfrac{1}{k}\\ x, & x\ge\dfrac{1}{k}\end{cases} \qquad (14)$$

This function is the modification of the three order spline function introduced by Yuan[5].

Performance analysis of smooth function: Before discussing the performance analysis, we need to introduce the following lemma:

Lemma 1: p(x,k) is defined as integral of sigmoid function (10) and x_+ is the plus function:

$\forall\rho>0, k\in R^+$:

$$p(x,k)^2-x_+^2\le\frac{(\log 2)^2}{k}+2\frac{\rho}{k}\log 2$$

The proof can be seen in[3].

Theorem 1: m(x,k) is defined as (14). For $x\in R, k\in R^+$

$$m(x,k)^2-x_+^2\le\frac{1}{24k^2}$$

The proof is omitted.

According to results of lemma1 and theorem 1, the following performance results of smooth functions (Theorem 2) are obtained.

Theorem 2: Let $\rho=\dfrac{1}{k}$

- The integral of sigmoid function (9), by lemma 1:

$$\begin{aligned}p(x,k)^2-x_+^2&\le\frac{(\log 2)^2}{k}+2\frac{\rho}{k}\log 2\\&=((\log 2)^2+2\log 2)\frac{1}{k^2}\\&\approx\frac{0.6927}{k^2}\end{aligned}$$

- The multiple knot spline function (14), by Theorem 1:

$$m(x,k)^2-x_+^2\le\frac{1}{24k^2}\approx\frac{0.0415}{k^2}$$

From Theorem 2, it is clear that m(x,k) is better than p(x,k). In order to show the difference more clearly, we present the following smooth performance comparison Fig. 1. The smooth parameter is set at k = 10.

As can be seen from Fig. 1, our proposed multiple knot spline function is closer to the plus function than sigmoid function, which indicates the superiority of our proposed smooth function.

MKS-SSVM model: If we replace the plus function in problem (13) by MKS function (13), a new smooth SVM model is obtained as following:

$$\frac{v}{2}\left\|m\left(e-D\left(K\left(A,A'\right)Du-e\gamma\right),\alpha\right)\right\|_2^2+\frac{1}{2}\left(u'u+\gamma^2\right) \qquad (15)$$

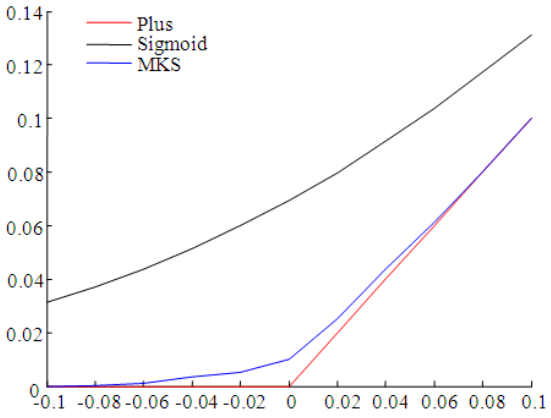

Fig. 1: Comparison figure between p(x,y) and m(x,k) at k = 10

It is called the Multiple Knot Spline Smooth Support Vector Machines (MKS-SSVM). We use Newton Armijo algorithm to solve the MKS-SSVM. The procedure of classification on dataset can be described as follows:

- Select an optimal parameter using uniform design[12] and 10-fold cross validation
- Solve MKS-SSVM using Newton Armijo algorithm
- Get separating plane
- Predict a new input
- Calculate accuracy of result

Application in diabetes disease diagnosis: We have used Pima Indian Dataset taken from UCI machine learning repository[6] in our applications. This dataset is commonly used among researchers who use machine learning method for diabetes disease classification, so it provides us to compare the performance of our method with that of others. The dataset contains 768 samples and two classes. The class distribution is:

Class 1: Normal (500)
Class 2: Pima Indian diabetes (268)

All samples have eight features. These features are:

- Number of time pregnant
- Plasma glucose concentration a 2 h in oral glucose tolerance test
- Diastolic blood pressure (mm Hg)
- Triceps skin fold thickness (mm)
- 2-h serum insulin ($\mu U\ mL^{-1}$)
- Body mass index (weight in kg/(height in m)^2)
- Diabetes pedigree function
- Age (years)

There has been a lot of research on medical diagnosis of diabetes disease in literature and most of them reported not too high classification accuracies. In Polat *et al*.[8] a cascade learning system based on Generalized Discriminant Analysis (GDA) and Least Square Support Vector Machien (LS-SVM) was used. They have reported 78.21% classification accuracy using LS-SVM with 10-fold cross validation (10×CV). They have also reported 79.16% classification accuracy using GDA-LS-SVM. Polat and Gunes[9] have reported 89.47% using Principal Component Analysis (PCA) and Adaptive Neuro-Fuzzy Inference System (ANFIS). The accuracy obtained by Kayaer and Yildirim[10] using General Regression Neural Network (GRNN) was 80.21%, while using Multilayer Neural Network (MLNN) with LM algorithm was 77.08%. Temurtas *et al*.[11] applied MLNN with LM and Probabilistic Neural Network (PNN) for diagnosing Pima Indian diabetes. They have reported 79.62% classification accuracy using MLNN with 10-fold CV and 82.37% accuracy with conventional (one training and one test) validation method. They have also reported 78.05% classification accuracy using PNN 10 fold CV and 78.13% accuracy using conventional validation method.

There have been several other studies reported with accuracy between 59.5 and 84.2%. The detail accuracy of these studies can be seen in Kahramanli, H and Allahverdi, N[14].

Parameter selection is one of the important steps in SSVM to improve classification accuracy. The performances of SSVM depend on the combination of several parameters. They are capacity parameter C, the kernel type K and its corresponding parameters. We used RBF kernel function, since of its good general performance and a few number of parameters[13] The parameters that should be optimized for the RBF kernel are the capacity parameter C and the kernel function parameter γ.

The 5-fold Cross Validation (CV) was used to select the best parameter. The data set is divided into 5 subsets and the holdout method is repeated 5 times. Each time, one of the 5 subsets is used as the test set and the other 4 subsets are put together to form a training set. The pairs of (C,γ) that the best CV accuracy is picked. After the (C,γ) is found, the whole training set is trained again to generate the final classifier. We used the nested Uniform Design (UD)[12] to choosing a good parameter.

RESULTS

To evaluate the effectiveness our method, we conducted experiments on Pima Indian diabetes dataset. All our experiments were performed on a personal computer, which utilizes a 2.00 GHz T7250 Intel(R) Core(TM) 2 duo CPU processor and 2550 megabytes of RAM. This computer runs on windows vista operating system, with MATLAB 7 installed. The classification accuracies obtained by the original SSVM and the new SSVM were presented in Table 1.

It can be seen from Table 1, the proposed MKS-SSVM can significant increase training accuracy and testing accuracy. However, the computational time of MKSSVM was not better than original SSVM.

Table 1: Classification accuracy between original SSVM and MKS-SSVM

	Original SSVM	MKSSVM
Best parameter (C,γ)	(1.78, 3.37e-005)	(316.23, 0.14)
Training accuracy (%)	77.66	93.24
Testing accuracy (%)	76.73	93.20
CPU time (sec)	399.762	700.1793

Table 2: Classification accuracies obtained with our method and other classifiers

Author	Method	Classification accuracy (%)
Kayaer and Yildrim[10]	GRNN	80.21
	MLNN with LM	77.08
Polat and Gunes[9]	PCA-ANFIS	89.47 (not reproducible)
Polat *et al*.[8]	LS-SVM	78.21
	GDA-LS-SVM	79.16
Temurtas, H. *et al*.[11]	MLNN with LM (10×FC)	79.62
	(10×FC)	78.05
	MLNN with LM (conventional valid)	82.37
	PNN (conventional valid)	78.13
Kahramanli, H and Allahverdi, N[14]	Various method	Between 59.5 and 84.2
This study	SSVM	76.73
	MKS-SSVM	93.20

For comparison purposes, Table 2 gives the classification accuracies of our method and previous methods.

Polat and Gunes[9] have reported 89.47% classification accuracy using PCA and ANFIS as seen in Table 2. Nevertheless, Temurtas[11] that used the same methods on Pima Indian diabetes obtained 66.78% classification accuracy. It is very far from 89.47% classification accuracy. On the other hand, the result of Polat and Gunes[9] using PCA-ANFIS methods on Pima Indian diabetes was not reproducible.

As we can see from Table 2, present method using MKS-SSVM obtained the highest classification accuracy so far.

DISCUSSION

A MKS-SSVM and its applications in diabetes disease diagnosis are presented. From theoretical aspect, MKS-SSVM has a better performance than the original SSVM. Likewise, when be applied to diagnosis of diabetes disease, the accuracy of this proposed method is better than original SSVM and previous studies.

Further exploration of the MKS-SSVM can yield more interesting results. We will apply this method for other classification problems.

CONCLUSION

This study has proposed new SSVM that called Multiple Knot Spline SSVM (MKS-SSVM). To evaluate the effectiveness our method, Pima Indian diabetes disease diagnosis was conducted. In order to achieve high classification accuracy, the Uniform Design approach was used to search the optimal MKS-SSVM parameters. The result was compared with the results of the original SSVM and the previous studies reported[8-11] focusing on Pima Indian diabetes diagnosis and using the same dataset.

As the conclusion, the following results can be summarized:

- The performance of MKS-SSVM was better than original SSVM
- It was seen that the MKS-SSVM obtained very promising result to help diagnosis of Pima Indian diabetes disease
- The classification accuracy of MKSSVM obtained by this study was better than the original SSVM and the previous studies reported

ACKNOWLEDGEMENT

This research was supported by GRS under grant no. 070147.

REFERENCES

1. Vapnik, V., 1995. The Nature of Statistical Learning Theory. Second edition, Springer-Verlag, New York. pp. 138-141. ISBN: 0-387-98780-0
2. Vapnik, V.1998. Statistical Learning Theory. Wiley, New York. pp. 493-496. ISBN: 0-471-03003-1
3. Lee, Y.J. and O.L. Mangasarian, 2001. A smooth support vector machine. J. Comput. Optimiz. Appli., 20: 5-22. DOI: 10.1023/A:1011215321374
4. Yuan, Y., J. Yan and C. Xu, 2005. Polynomial Smooth Support Vector Machine (PSSVM). Lectures Notes in Computer Science, Springer, DOI: 10.1007/11527503_19
5. Yuan, Y., W. Fan and D. Pu, 2007. Spline function smooth support vector machine for classification. J. Ind. Manage. Optimiz., 3: 529-542. http://aimsciences.org/journals/redirecting.jsp?paperID=2609
6. Newman, D.J., Hettich, S., Blake, C. L. S., & Merz, C. J., 1998. UCI repository of machine learning database, Irvine, CA: University of California, Dept. of Information and Computer Science. http://www.ics.uci.edu/~mlearn/~MLRepository.html

7. Mangasarian, O.L.,2000. Generalized Support Vector Machines. In: Advances in large Margin Classifiers, Smola, A., P. Bartlett, B. Scholkopf and D. Schurrmans (Eds.). MIT Press, Cambridge, MA.,ISBN: 0-262-19448-1, pp: 35-146.
8. Polat, K., S. Gunes and A. Aslan, 2008. A Cascade Learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. Expert Syst. Appli., 34: 214-221. DOI: 10.1016/j.eswa.2006.09.012
9. Polat, K. and S. Gunes, 2007. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Digit. Sign. Proc., 17: 702-710. DOI: 10.1016/j.dsp.2006.09.005
10. Kayaer, K.and T. Yildirim, 2003. Medical diagnosis on Pima Indian diabetes using general regression neural networks. Proceedings of the International Conference on Artificial Neural Networks and Neural Information Processing, Springer, Istanbul-Turkey, June 26-29, pp: 181-184. http://www.yildiz.edu.tr/~tulay/publications/Icann-Iconip2003-2.pdf
11. Temurtas, H. *et al*., 2009. A comparative study on diabetes disease using neural networks. Expert Syst. Appli., 36: 8610-8615. DOI: 10.1016/j.eswa.2008.10.032
12. Huang, C.M., Y.J. Lee, D.K.J. Lin and S.Y. Huang, 2007. Model selection for support vector machines via uniform design. Comput. Stat. Data Anal., 52: 335-346. http://EconPapers.repec.org/RePEc:eee:csdana:v:52:y:2007:i:1:p:335-346
13. Hsu, C.W., C.C. Chang and C.J. Lin, 2003. Practical guide to support vector classification. Department of Computer Science and Information Engineering National Taiwan University. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf
14. Kahramanli, H and Allahverdi, N, 2008. Design of a hybrid system for the diabetes and heart diseases. Expert Syst. Appli., 35: 82-89. DOI: 10.1016/j.eswa.2007.06.004