Journal of Computer Sciences and Applications, 2015, Vol. 3, No. 6, 181-184
Available online at http://pubs.sciepub.com/jcsa/3/6/14

DOI:10.12691/jcsa-3-6-14

Comparison of Kernel Selection for Support Vector Machines Using Diabetes Dataset

Tapas Ranjan Baitharu[1], Subhendu Ku. Pani[1,*], Sunil kumar Dhal[2]

[1]Department of Computer Science & Engineering, Orissa Engineering College, Bhubaneswar, Odisha (India) under Biju Patnaik University of Technology, Odisha
[2]Department of Computer Science & Engineering, SriSri University, Bhubaneswar, Odisha, India
*Corresponding author: skpani.india@gmail.com

Abstract One of the major problems in the study of Support vector machine (SVM) is kernel selection, that's based necessarily on the problem of deciding a kernel function for a particular task and dataset. By contradiction to other machine learning algorithms, SVM focuses on maximizing the generalisation ability, which depends on the empirical risk and the complexity of the machine. We were focused on SVM trained using linear, polynomial, puk and Radial Basic Function (RBF) kernels. A preliminary study has been made between SVM using the best choice of kernel. Results had revealed that SVM trained using Linear Kernel is the best choice for dealing with Diabetes dataset.

Keywords: *data mining, machine learning algorithms, support vector machine, kernels function*

Cite This Article: Tapas Ranjan Baitharu, Subhendu Ku. Pani, and Sunil kumar Dhal, "Comparison of Kernel Selection for Support Vector Machines Using Diabetes Dataset." *Journal of Computer Sciences and Applications*, vol. 3, no. 6 (2015): 181-184. doi: 10.12691/jcsa-3-6-14.

1. Introduction

Data mining. In different kinds of information databases, such as scientific data, medical data, financial data, and marketing transaction data; analysis and finding critical hidden information has been a focused area for researchers of data mining [1,2,4]. How to effectively analyze and apply these data and find the critical hidden information from these databases, data mining technique has been the most widely discussed and frequently applied tool from recent decades. Although the data mining has been successfully applied in the areas of scientific analysis, business application, and medical research and its computational efficiency and accuracy are also improving, still manual works are required to complete the process of extraction [13].

Support vector machine (SVM) was the first proposed kernel based algorithm. It uses a kernel function to transform data from input space into a high dimensional feature space in which it searches for a separating hyperplane [3]. Linear, polynomial, RBF and others kernel functions are commonly used to transform input space into desired feature space. Furthermore, SVM is based on the principle of Structure Risk Minimization by taking into account of the probability of misclassifying yet to be seen patterns for a fixed but unknown probability distribution of data. It uses a linear separating hyperplane to create a classifier, yet it is not easy to separate some problems in the original input space linearly. But it can easily transform the original input space into a high dimensional feature space nonlinearly, where it is trivial to find an optimal linear separating hyperplane. In this paper, we were attempted to investigate the best choice among SVM kernels namely linear, puk, polynomial and RBF [14].

This paper was organized as follow: in section 2 techniques and algorithms is briefly described. In section 3 experimental study and analysis is presented. Finally the conclusion is discussed in section 4.

2. Techniques and Algorithms

Researchers identify two fundamental goals of data mining: prediction and description. Prediction makes use of existing variables in the database in order to predict unknown or future values of interest, while description focuses on finding patterns describing the data the subsequent presentation for user interpretation. The relative emphasis of both prediction and description differs with respect to the underlying application and technique. There are several data mining techniques fulfilling these objectives.

2.1. Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis [2]. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification [2]. In Learning the training data are analyzed by classification algorithm. In classification test

data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples [5,6]. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these preclassified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier. Some well known classification models are: a) Classification by decision tree induction b) Bayesian Classification c) Neural Networks d) Support Vector Machines (SVM) [7,10].

2.2. Clustering

Clustering is a technique for identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality. Some commonly used clustering methods are: a) Partitioning Methods b) Hierarchical Agglomerative (divisive) methods c) Density based methods d) Grid-based methods e) Model-based methods

2.3. Association Rules

An Association Rule is a rule of the form milk and bread =>butter,where 'milk and bread' is called the rule body and butter the head of the rule. It associates the rule body with its head.In context of retail sales data, our example expresses the fact that people who are buying milk and bread are likely to buy butter too. This association rule makes no assertion about people who are not buying milk or bread.We now define an association rule:Let D be a database consisting of one table over n attributes {a1, a2, . . . , an}. Let this table contain k instances.The attributes values of each ai are nominal1. In many real world applications (such as the retail sales data) the attribute values are even binary (presence or absence of one item in a particular market basket) [8,9]. In the following an attribute-value-pair will be called an item. An item set is a set of distinct attribute-valuepairs. Let d be a database record. d satisfies an item set X _ {a1, a2, . . . , an} if X _ d.An association rule is an implication X) Y where X, Y _ {a1, a2, . . . , an}, Y 6= ; and X \ Y = ;. The support s(X) of an item set X is the number of database records d which satisfy X.Therefore the support s(X) Y) of an association rule is the number of database records that satisfy both the rule body X and the rule head Y. Note that we define the support as the number of database records satisfying X \ Y, in many papers the support is defined as s(X\Y) k. They refer to our definition of support as support count. The confidence ˆc(X) Y of an association rule X) Y is the fraction ˆc(X) Y) = s(X\Y) s(X). From a logical point of view the body X is a conjunction of distinct attribute-value-pairs.

3. Experimental Study and Analysis

3.1. WEKA Tool

We use WEKA (www.cs.waikato.ac.nz/ml/weka/), an open source data mining tool for our experiment. WEKA is developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state-of-the-art tool for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data pre-processing, feature reduction, classification, regression, clustering, and association rules. It also includes visualization tools. The new machine learning algorithms can be used with it and existing algorithms can also be extended with this tool.We performed computer simulation on a Diabetes dataset available UCI Machine Learning Repository [11,15].

3.2. Results Analysis

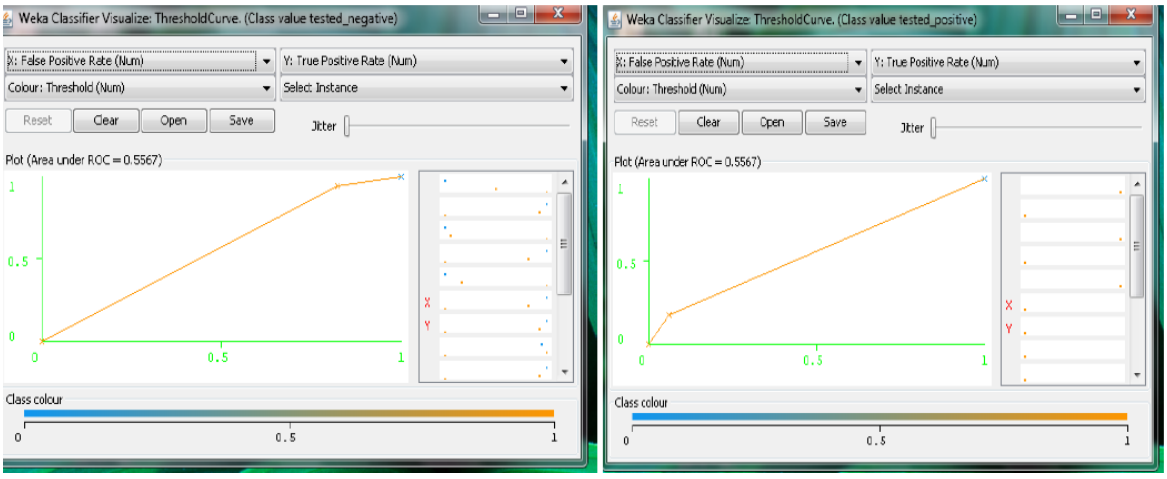

Figure 1. Visualize ThresholdCurve using Normalized Poly Kernel of SMO(Class Value:Tested Negative and Tested Positive)

We compare different kernel of SVM and its performance is discussed in Table 1 [12]. Figure 5 shows the accuracy result. Visualize ThresholdCurve using Normalized Poly Kernel of SMO(Class Value: Tested

Negative and Tested Positive) is shown in Figure 1. Visualize ThresholdCurve using RBF Kernel of SMO (Class Value:Tested Negative and Tested Positive) is shown in Figure 2. Visualize ThresholdCurve using Linear Kernel of SMO(Class Value: Tested Negative and Tested Positive) is shown in Figure 3. Visualize ThresholdCurve using Puk Kernel of SMO(Class Value: Tested Negative and Tested Positive) is shown in Figure 4.

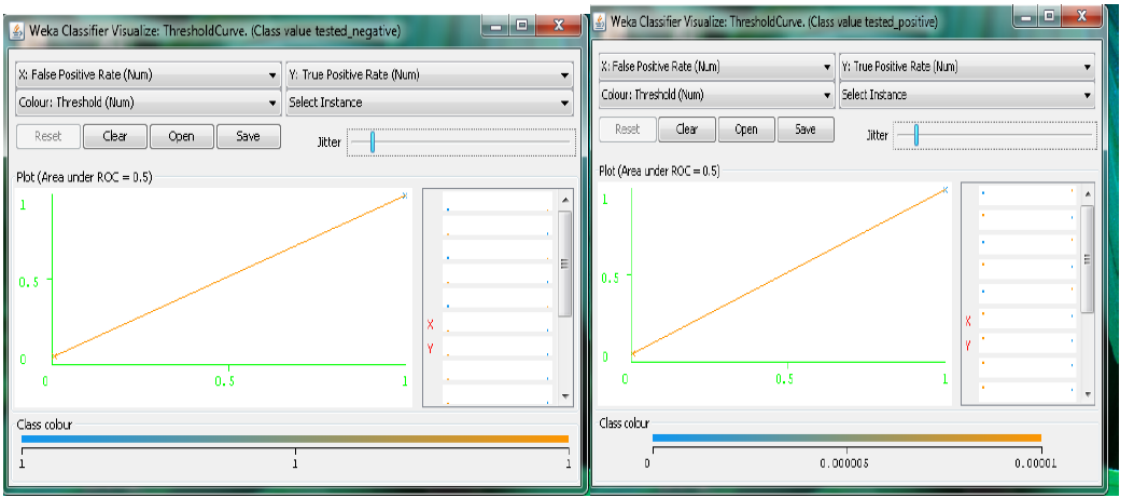

Figure 2. Visualize Threshold Curve using RBF Kernel of SMO (Class Value: Tested Negative and Tested Positive)

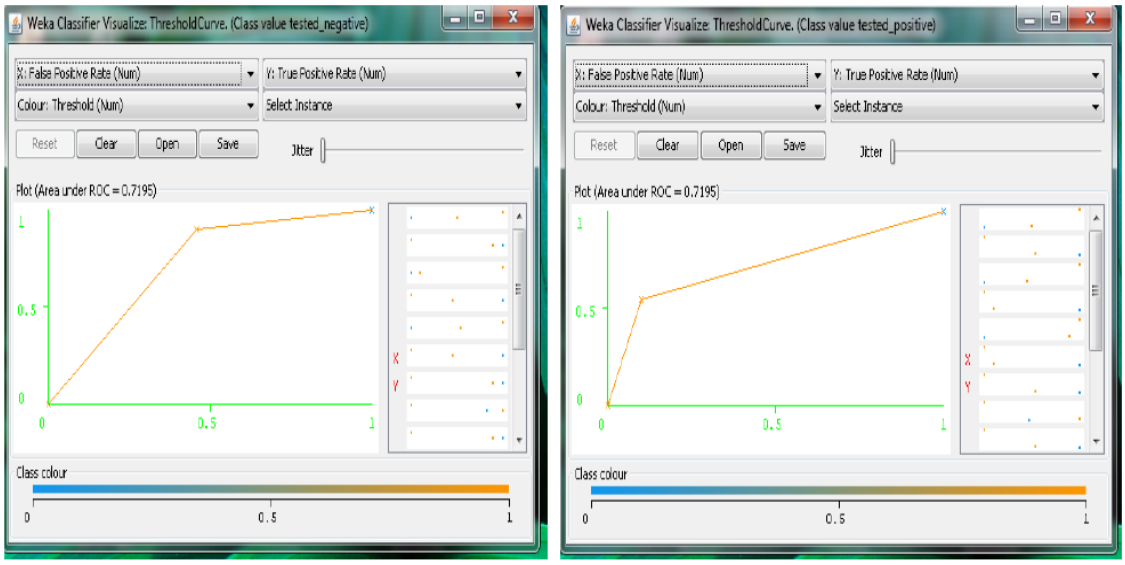

Figure 3. Visualize Threshold Curve using Linear Kernel of SMO(Class Value: Tested Negative and Tested Positive)

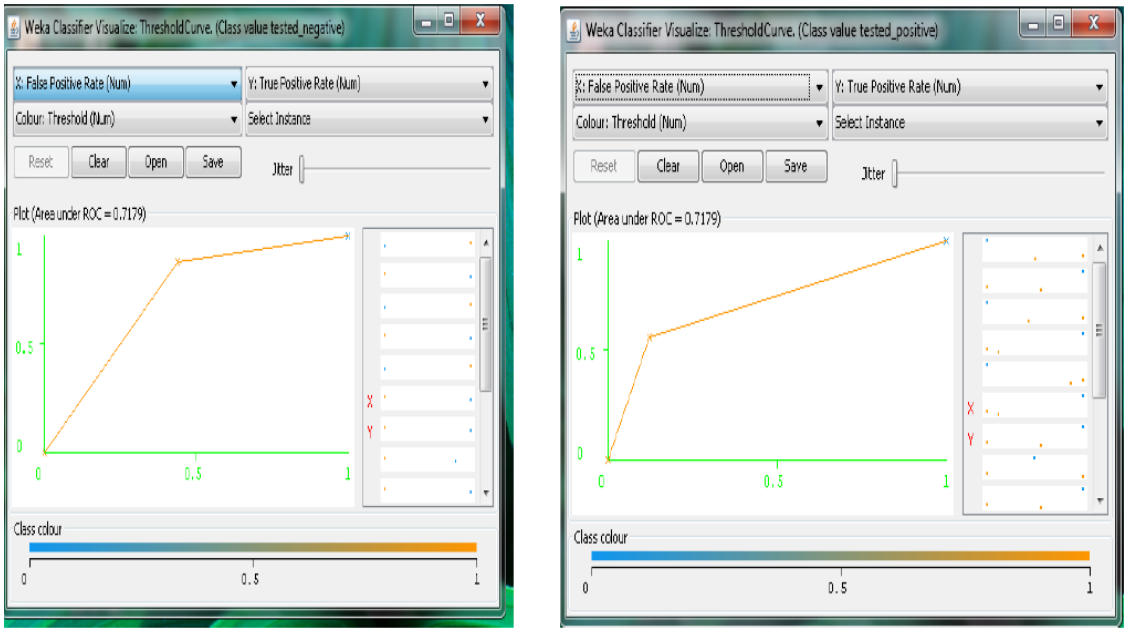

Figure 4. Visualize Threshold Curve using Puk Kernel of SMO(Class Value: Tested Negative and Tested Positive)

Table 1. Performance analysis of different Kernels of SVM

Name of the Kernel	Number of suport Vectors	Number of kernel Evaluation	Time Taken to Build the Model	Root Mean Squared Error	Accuracy(%)
Normalized Poly Kernel	544	281044	0.98	0.5728	67.1875
RBF Kernel	545	282572	0.87	0.5907	65.1042
Linear Kernel		19131	0.03	0.476	77.3438
Puk Kernel	439	260754	0.33	0.4841	76.5625

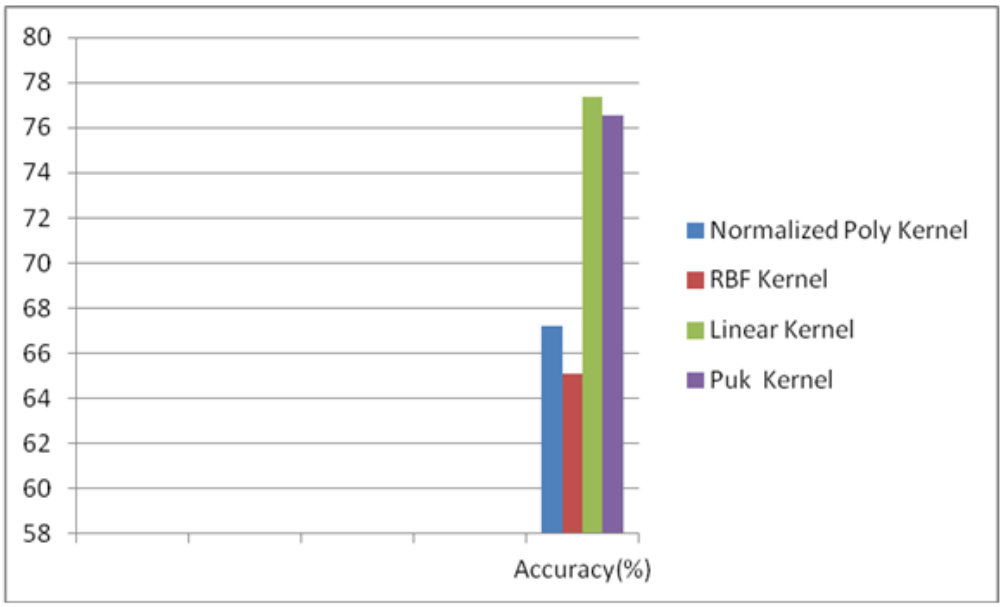

Figure 5. Accuracy Comaprision of different Kernels of SVM

4. Conclusion

In this paper, we have presented the performance of SVM algorithm using different kernels on diabetes dataset and a comparison was made. In deed, we were attempted to investigate the best choice among SVM kernels namely linear, polynomial, puk and radial basis function (RBF) kernels. Different degree of the polynomial kernel and different width of the RBF kernel were evaluated. Our studies reveal that SVM linear kernel provided the best performance than other kernels.

References

[1] Klosgen W and Zytkow J M (eds.), Handbook of data mining and knowledge discovery, OUP, Oxford, 2002.

[2] Provost, F., & Fawcett, T., Robust Classification for Imprecise Environments. Machine Learning,Vol.42, No.3, pp. 203-231,2001.

[3] Larose D T, Discovering knowledge in data: an introduction to data mining, John Wiley, New York, 2005.

[4] Kantardzic M, Data mining: concepts, models, methods, and algorithms, John Wiley, New Jersey, 2003.

[5] Goldschmidt P S, Compliance monitoring for anomaly detection, Patent no. US 6983266 B1, issue date January3, 2006, Available at: www.freepatentsonline.com/6983266.html.

[6] Bace R, Intrusion Detection, Macmillan Technical Publishing, 2000.

[7] A. B. M. S. Ali, and A. Abraham, "An Empirical Comparison of Kernel Selection for Support Vector Machines", Soft computing systems: design, management and applications, Ajith Abraham, Javier Ruiz-del-Solar,Mario Köppen (eds.), IOS Press Publisher, Amsterdam, 2002, pp. 321-330.

[8] Agrawal R. and Srikant R. Fast Algorithms for Mining Association Rules. In M. Jarke J. Bocca and C. Zaniolo, editors, Proceeedings of the 20th International Conference on Very Large Data Bases (VLDB'94), pages 475-486, Santiago de Chile, Chile, Sept 1994. Morgan Kaufmann.

[9] Scheffer T. Finding Association Rules That Trade Support Optimally against Confidence. Unpublished manuscript.

[10] M. Bak, "Support Vector Classifier with Linguistic Interpretation of the Kernel Matrix in Speaker Verification", Man-Machine Interactions, Krzysztof A. Cyran, Stanislaw Kozielski, James F. Peters (eds.), ISSN 1867-5662, Springer, 2009, pp 399-406.

[11] UCI Machine Learning Repository, Available at http://archive.ics.uci.edu/ml/machine-learningdatabases/statlog/german/.

[12] C. J. C. Burges and B. Schölkopf, "Improving the Accuracy and Speed of Support Vector Learning Machines", Advances in Neural Information Processing Systems 9, Cambridge, MIT Press, 1997, pp. 375-381.

[13] Berry M J A and Linoff G S, Data mining techniques: for marketing, sales, and relationship management, 2nded (John Wiley; New York), 2004.

[14] M. Bentoumi, G. Millerioux, G. Bloch, L. Oukhellou and P. Aknin, "Classification de Défauts de Rail par SVM," Congrès International IEEE de Signaux, Circuits et Systèmes SCS'04, Tunisie, 2004, pp. 242-245.

[15] Fuchs G, Data Mining: if only it really were about Beer and Diapers, Information Management Online, July 1, (2004), Available at: http://www.informationmanagement. com/ news/10061331.html.