JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

STUDY OF PERFORMANCE OF SVM ALGORITHM BASED ON PEARSON UNIVERSAL KERNEL FOR DATA CLASSIFICATION

[1]Karthik Sundararajan, [2]Manjushree K

[1]Assistant Professor, [2]Assistant Professor
[1,2] Department of Computer Science & Engineering
[1.2] B.N.M. Institute of Technology, Bengaluru.

Abstract: Big data mining can be referred as the methodology of inferring meaningful information from large datasets. The type of algorithms used for mining information from such data influences the efficiency and performance of the overall process. Classification is one of the well-known problems under the domain of data mining. It attempts to classify the data into multiple groups. There are variety of data mining algorithms available for classification such as naïve Bayesian, Fuzzy logic, Latent Dirichlet Allocation, Decision trees etc. One such model which possess reasonably better accuracy for majority of the cases is the Support Vector Machines. The reason for the success of SVM lies in the fact that users can choose an appropriate kernel function according to the complexity of relationship model. In this work, Pearson VII Universal Kernel (PUK) function based Support vector machine is implemented for chosen datasets and its performance compared with various other classification models. Experimental results show that for the completely two different datasets that are used in this work, PUK based SVM outperforms all other classification algorithms with an accuracy of 100% followed by RBF based SVM achieving 100% and 99.51% respectively for breast cancer dataset and bank information dataset.

***IndexTerms* – Data mining, SVM, Pearson VII kernel, Classification**

I. INTRODUCTION

Big data [1] can be defined as massive amount of datasets that are difficult to be handled by the existing algorithms and technologies. Big data are usually characterized by four V's such as Volume, Velocity, Variety, Veracity. In recent times literature speaks additionally about Value and Visualization. Mining information from huge datasets is a challenging task. Various mining activities can be performed in data such as Clustering, Classification, Finding out association rules and Sequence patterns. The ultimate objective of big data mining [2] is arriving at a meaningful summarization which will make it easier to understand the nature of data. Figure 1 depicts the steps involved in data mining.

Classification partitions the given data or objects into various classes. There are three major categories of classification algorithms. They are Supervised learning, Unsupervised learning and Reinforcement Learning. In Supervised learning the algorithm works on labeled data. In other words, both input and output are known beforehand. The objective will be to learn a function from set of known inputs and outputs. Classification will come under supervised learning. Models such as Discriminant analysis, Naïve bayes, Support vector machines, Artificial neural networks, Decision trees etc., will come under classification algorithms. Unsupervised learning algorithms works on unlabeled data. The user is provided with only the input and the class assignment has to be done without knowing the output classes in advance. Clustering is a well-known unsupervised learning technique. Clustering algorithms can further be divided into partitioning algorithms, Hierarchical algorithms and spectral clustering algorithms.

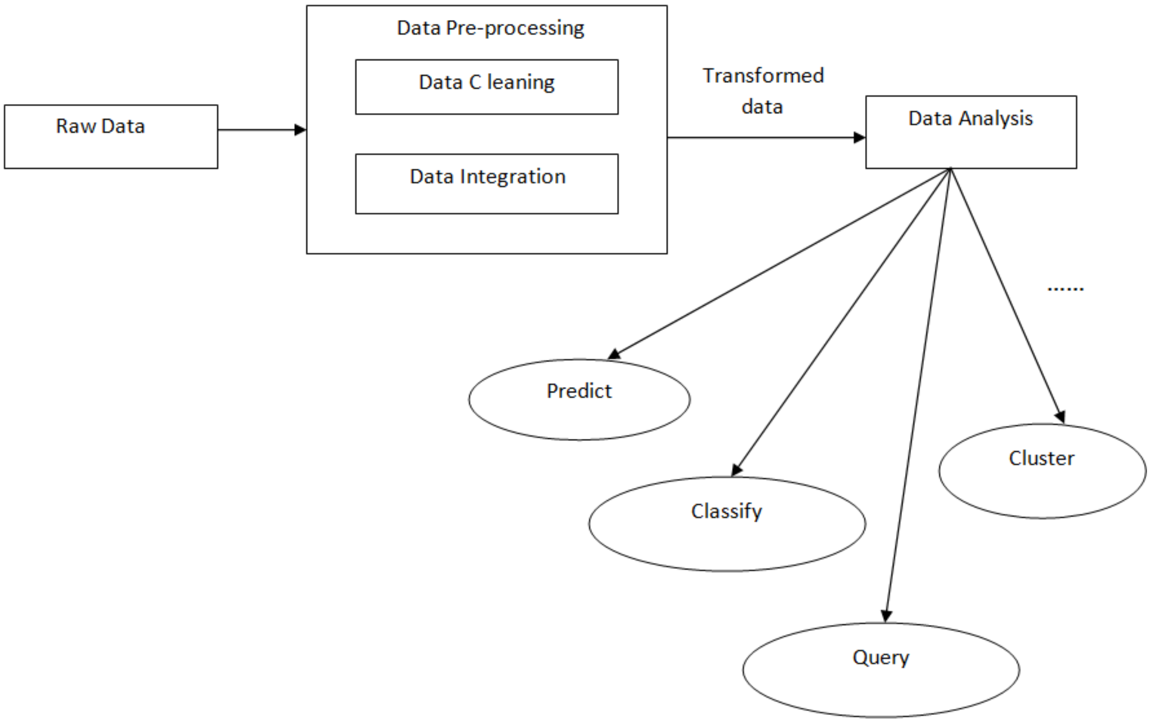

Figure 1: Data mining process

II. PROPOSED WORK

In this work, Pearson VII Universal Kernel function based SVM is implemented for bank datasets and Breast cancer datasets and is compared with the performance of other algorithms such as Multi-layer perceptron, Logistic Regression, Stochastic gradient Descent algorithms. The results are depicted and tabulated in the experimental results section. In the following section a brief analysis of SVM is given.

Support Vector Machine

Support vector machines(SVMs)[10] was initially proposed by vapnik as a novel solution to classification problems. It is a learning algorithm under supervised learning techniques that classifies objects into one of the two classes, if it is a two-class SVM, or classifies objects into one of the many classes, if it is a multi-class SVM. SVM converts the data into higher dimensions using non-linear mapping. The objective of a SVM is to build a learning function that accurately predicts the class to which an object belongs.

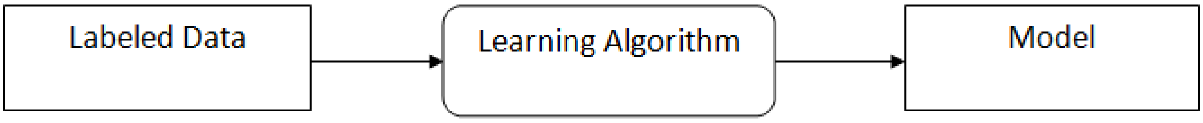

Figure. 2: Supervised learning process

Let $\{(x_1,y_1), (x_2,y_2), (x_3,y_3),\ldots,(x_n,y_n)\}$ denote the set of training data given as input to the classifier.
Here $x=(x_1,x_2,..,x_n)$ represents the input vector and y_i is the corresponding class label for the input data. $y_i \in [0,1]$ where 0 is considered as negative class and 1 is considered as positive class. The linear function [3], [4] of SVM is of the form

$$f(x) = w.x + b \qquad (1)$$

where w is the weight vector and b is the bias. The input vector will be assigned its class based on the value of f(x).

$y_i = 0$, if $f(x)>=0$

$y_i =1$, if $f(x)<0$

The equation of a hyperplane can be derived from equation 1 and is denoted as follows:

$$w.x + b =0 \qquad (2)$$

The hyperplane divides the input data space into two halves as in positive and negative. In a 2-Dimensional space a hyperplane is simply viewed as a line and in 3-Dimension it is seen as a plane.

Linear Separability

In the input data space, if the positive and negative samples of data can be separated then it can be called as Linearly Separable case and if they cannot be separated then it comes under linearly non-separable case. The separation line which distinguishes the positive and negative samples is called as hyperplane or Decision rule in the SVM classifier.

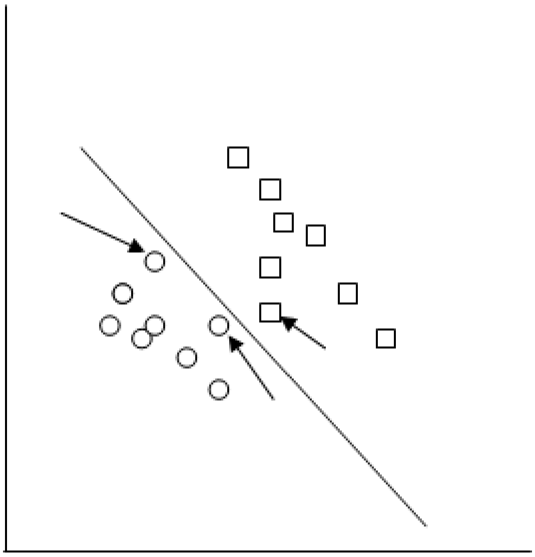

Figure 3a: Linearly Separable

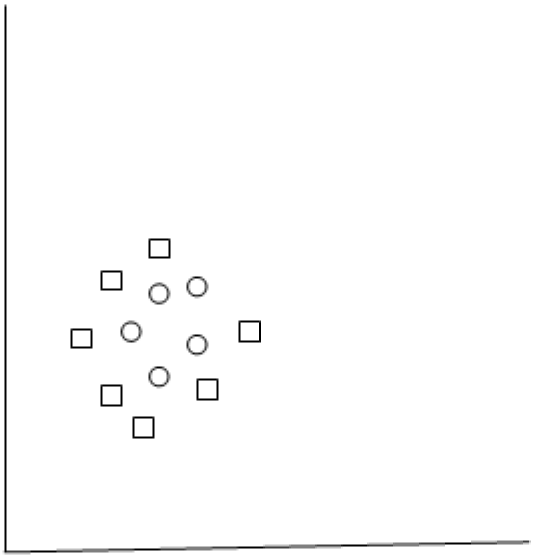

Figure 3b: Linearly Non-separable

Figure 3a and 3b represents sample data distribution of linearly separable and inseparable cases. The circles in the figure represents the sample data points belonging to the positive class. The small squares denote the data points that belongs to a negative class. Figure 3a depicts a linearly separable case where both positive and negative points are separated by a hyperplane. In figure 3b it can be seen that the positive points and negative points cannot be separated by any line. Hence such cases belong to linearly non-separable cases. The small arrows in figure 3a represents the support vectors for the given set of data points. The points which are closer to the hyperplane will be denoted as support vectors as it is assumed to have maximum effect on the hyperplane.

Kernel Functions

Kernel functions [6], [7] are used in SVMs to perform computations in higher dimensional space. It computes dot products. The function f maps data from n-dimensional space to m-dimension.

$$K(x,y)= <f(x).f(y)> \qquad (3)$$

There are various kernel functions available in the literature. The frequently used ones in the literatures are discussed below.

(i)Linear Kernel Function

The equation for linear kernel function is given as the sum of the support vectors. With the support vectors, hyperplane with maximum margin will be computed.

$$K(x_i,x_j)= <xi.xj> + k \qquad (4)$$

Where k is a constant and K denotes the kernel function.

(ii)Polynomial Kernel Function

The polynomial kernel function is similar to the linear kernel function but with a degree, d. It can be formulated as

$$K(x_i,x_j)= (<xi.xj> + k)^{d} \qquad (5)$$

(iii)Gaussian Radial Basis Function

Another most widely used kernel is the gaussian radial basis kernel function. The kernel function is defined as

$$K(x_i,x_j)= \exp(-\left\| x_i , x_j \right\|^{2} / 2\sigma^{2}) \qquad (6)$$

σ is the parameter that control the width of the gaussian.

(iv)Pearson VII Universal kernel Function

PUK function[8], [9] is one of the latest kernel functions used in Support vector machines. The general curve fitting equation of PUK function can be defined as

$$f(x) = H / [1 + (2(x_i-x_j) \sqrt{2^{\left(\frac{1}{\omega}\right)} - 1/}\ \sigma)^{2}]^{\omega} \qquad (7)$$

where H is the peak and w, σ are the parameters that control the half- width of the input space. This Pearson function is used as a kernel function of SVM in this work for the purpose of classification.

III. RESULTS AND DISCUSSIONS

The datasets used in this study are Breast cancer Wisconsin dataset from UCL machine learning repository and a bank information dataset. The breast cancer dataset consists of 11 attributes and 699 records of clinical cases during the period of 3 years at 8 different months. The first 10 attributes represent the id_number, clump thickness, Uniformity of cell size, uniformity of cell shape, Marginal adhesion, Single epithelial cell size, bare nuclei, normal nucleoli, Mitoses and the last attribute is the class attribute with yes and no values, denoting benign and malignant. The bank dataset contains 17 attributes and 4521 records about various customers and their information and also information relating that person with the bank. Implementation of this work is done using weka 3.8 software program and algorithms such as SVM, MLP, Logistic Regression and Stochastic Gradient Descent were implemented in weka software program.

Table 1: Tabulated result of Breast Cancer dataset

	MLP	**Logistic Regression**	**Stochastic Gradient Descent**	**Support Vector Machine**		
				PolyKernel	**RBFkernel**	**PUK kernel**
Accuracy (%)	99.14	96.70	97.13	96.99	100	**100**
Kappa statistic	0.9811	0.9271	0.9369	0.9337	1	**1**
RMS error	0.0939	0.1528	0.1692	0.1773	0	**0**
Mean absolute error	0.0192	0.0455	0.0286	0.03	0	**0**

Table 2: Detailed Class-wise accuracy of Breast Cancer dataset

		MLP	**Logistic Regression**	**Stochastic Gradient Descent**	**Support Vector Machine**		
	Class				**PolyKernel**	**RBFkernel**	**PUK kernel**
TP rate	Yes	0.996	0.950	0.967	0.963	1	**1**
	No	0.989	0.976	0.974	0.974	1	**1**
FP rate	Yes	0.011	0.024	0.026	0.026	0	**0**
	No	0.004	0.050	0.033	0.037	0	**0**
Precision(%)	Yes	98	95.4	95.1	95.1	100	**100**
	No	99.8	97.4	98.2	98	100	**100**
Recall(%)	Yes	99.6	95	96.7	96.3	100	**100**
	No	98.9	97.6	97.4	97.4	100	**100**
F-Measure(%)	Yes	98.8	95.2	95.9	95.7	100	**100**
	No	99.3	97.5	97.8	97.7	100	**100**

Table 1 tabulates the accuracy results of Multi-layer perceptron, Logistic regression, Stochastic Gradient Descent and SVM algorithms. From the table it is clear that best accuracy is obtained for SVM based classification. Though SVM attains best accuracy, it is not the case for all kernel functions [5]. As we can see compared to PolyKernel, RBF kernel and PUK kernel does better and achieves 100% accuracy whereas PolyKernel function achieves 96.99% accuracy.

The Kappa statistic measure is used to signify the agreement of the predicted result with the actual class. The Value of 1 signifies perfect agreement and the value 0 denotes complete disagreement. The root mean squared error measures the quadratic average magnitude of the error. Mean absolute error calculates the average magnitude of errors. In other words, accuracy is measured for continuous objects. TP rate is the true positives rate. It can be defined as the number of samples that are predicted positive being actually positive. FP rate is the false positives rate. It can be defined as the number of samples which are predicted positive but are actually negative

Precision [9] can be defined as the measure determines how much fraction of the samples predicted positive are positive

Precision = True Positive / (True Positive + False Positive) (8)

Recall can be defined as the measure that determines how much fraction of samples that are positive were actually predicted positive

Recall = True Positive / (True Positive + False Negative) (9)

F-Measure is defined as the weighted average of precision and recall. It can be represented as

F-Measure = 2 * (precision * recall) / (precision + recall) (10)

Table 3: Tabulated result of Bank information dataset

	MLP	Logistic Regression	Stochastic Gradient Descent	Support Vector Machine		
				PolyKernel	**RBFkernel**	**PUK kernel**
Accuracy (%)	97.45	90.51	89.38	89.29	99.51	**100**
Kappa statistic	0.8706	0.4076	0.2557	0.2197	0.9757	**1**
RMS error	0.1507	0.2668	0.3288	0.3272	0.0698	**0**
Mean absolute error	0.0326	0.1346	0.1062	0.1071	0.0049	**0**

Table 4: Detailed Class-wise accuracy of Bank information dataset

		MLP	Logistic Regression	Stochastic Gradient Descent	Support Vector Machine		
	Class				**PolyKernel**	**RBFkernel**	**PUK kernel**
TP rate	Yes	0.848	0.342	0.194	0.159	0.958	**1**
	No	0.991	0.979	0.985	0.989	1	**1**
FP rate	Yes	0.009	0.022	0.015	0.012	0.00	**0**
	No	0.152	0.658	0.806	0.841	0.042	**0**
Precision	Yes	92	67.4	62.7	64.3	100	**100**
	No	98	91.9	90.4	90.0	99.5	**100**
Recall	Yes	84.8	34.2	19.4	15.9	95.8	**100**
	No	99.1	97.9	98.5	98.9	100	**100**
F-Measure	Yes	88.5	45.4	29.6	25.5	97.8	**100**
	No	98.6	94.8	94.3	94.2	99.7	**100**

Table 3 and 4 depicts the performance of the SVM and other classification algorithms on the bank information dataset. From the result of table 3 it is clear that Pearson VII Universal kernel function based SVM outperforms all other classification models as well as the SVM implemented through PolyKernel and RBF kernel function. Unlike Breast cancer dataset where RBF kernel produced an 100% accuracy, in bank information dataset PUK function based SVM gives an 100% accuracy whereas RBF kernel based SVM gives about 99.51 % accuracy. And also from table 1 and table 3 it can be noted that Root mean square error for RBF kernel and PUK kernel is minimum but it is higher for PolyKernel function in both the datasets. Hence it can be safely assumed that out of the three kernel functions, PolyKernel function has the least performance. RBF kernel performs equal to PUK function in breast cancer dataset but gives a slightly lesser accuracy in bank dataset when compared to PUK function based SVM.

IV. CONCLUSION

Big data classification is one of challenging data mining problems. In this work classification is performed for breast cancer datasets and bank information datasets. The performance of Pearson VII Universal Kernel function based SVM is compared with various other kernel function such as PolyKernel and RBF kernel. Along with that the performance of PUK based SVM is compared with other classification models such as Multi-layer perceptron, Logistic Regression and SGD methods. From the experimental results it can be concluded that PUK function based SVM achieves the best classification accuracy of 100% for both datasets followed by Gaussian RBF kernel function based SVM which achieves 100% and 99.51% accuracy for the respective datasets. PolyKernel function based SVM attains the least performance when compared to other kernel functions of SVM.

REFERENCES

[1] Chen, Min, Shiwen Mao, and Yunhao Liu. "Big data: a survey." *Mobile Networks and Applications* 19.2 (2014): 171-209.

[2] Fan, Wei, and Albert Bifet. "Mining big data: current status, and forecast to the future." *ACM sIGKDD Explorations Newsletter* 14.2 (2013): 1-5.

[3] Wu, Jianxin, and Hao Yang. "Linear regression-based efficient svm learning for large-scale classification." *IEEE transactions on neural networks and learning systems* 26.10 (2015): 2357-2369.

[4] Patra, Swarnajyoti, and Lorenzo Bruzzone. "A novel SOM-SVM-Based active learning technique for remote sensing image classification." *IEEE Transactions on Geoscience and Remote Sensing* 52.11 (2014): 6899-6910.

[5] Liu, Lan, et al. "Thematic information detection for remote sensing image using SVM kernel functions." *Signal Processing, Communications and Computing (ICSPCC), 2015 IEEE International Conference on*. IEEE, 2015.

[6] Patle, Arti, and Deepak Singh Chouhan. "SVM kernel functions for classification." *Advances in Technology and Engineering (ICATE), 2013 International Conference on*. IEEE, 2013.

[7] Hussain, Muhammad, et al. "A comparison of SVM kernel functions for breast cancer detection." *Computer Graphics, Imaging and Visualization (CGIV), 2011 Eighth International Conference on*. IEEE, 2011.

[8] Abakar, Khalid AA, and Chongwen Yua. "Performance of SVM based on PUK kernel in comparison to SVM based on RBF kernel in prediction of yarn tenacity." *Indian Journal of Fibre & Textile Research* 39 (2014): 55-59.

[9] Zhang, Guangya, and Huihua Ge. "Support vector machine with a Pearson VII function kernel for discriminating halophilic and non-halophilic proteins." *Computational biology and chemistry* 46 (2013): 16-22.

[10] Liu, Bing. *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media, 2007.