Open access Journal International journal of Emerging Trends in Science and Technology

Investigation of the effect of Training Data on the Performance of Support Vector Machine in Classification of BrainMR Images

Authors

Swetha K. T[1],Sharath Kumar K. T[2], Sanath Kumar M. T[3], Basavaraj L[4]

[1]Department of Digital Electronics & Communication Systems, VTU Regional College
Mysore, India.
Email:*2007.swetha@gmail.com*
[2]Department of Telecommunications, RV College of Engineering
Bengaluru, India.
Email:*yuvan8888@gmail.com*
[3]Department of Digital Electronics & Communication Systems, VTU Regional College
Mysore, India
Email:*mtsanathkumar@gmail.com*
[4]Department of Electronics &Communication Engineering, ATME College of Engineering
Mysore, India
Email:*basavaraj.atme@gmail.com*

Abstract

In recent years, a wide research is being carried out on brain imaging which involves computer aided detection of abnormalities in brain. Out of many diagnostic imaging techniques for the early detection of any abnormal changes in brain tissues, Magnetic Resonance Imaging (MRI) is a widely-used imaging method. The shortage of radiologists for analyzing the brain MR images calls for an automated system to analyze and classify such medical images. Support Vector Machine (SVM) has been widely used in the recent years to classify brain MR images into different classes. SVM Classifiers perform the task of classification in two phases – training phase and testing phase. The amount of image data to be used for training plays a vital role in determining the accuracy of the SVM. This paper focuses on determining the optimal number of image data in the training set for which a better classification accuracy is obtained. Classification experiments with various percentages of data in the training set show that 80% of total dataset is the optimal one. Results also point out that Polynomial kernel function of SVM is more apt for brain MR images classification with classification accuracy of 100% when trained with 80% of data.

Keywords:GLCM, Median filtering, SVM kernels, Texture features, Training dataset

INTRODUCTION

Magnetic Resonance Imaging (MRI) has become a widely-used method of high quality medical imaging, especially in brain imaging where soft tissue contrast and non-invasiveness are clear advantages. MR images are examined by radiologists based on visual interpretation of the films to identify the presence of tumour abnormal tissue. The shortage of radiologists and the large volume of MRI to be analyzed make such readings labor intensive, cost expensive and often inaccurate. The sensitivity of the human eye in interpreting large numbers of images decreases with increasing number of cases, particularly when only a small number of slices are affected. Also, efficient diagnosis in short period of time is the need of the hour. This calls for an automated system to analyze and classify such medical images.

In the last few years there has been growing interest in the use of machine learning classifiers for analyzing MRI data. A classifier is a function that takes the values of various features in an example and predicts the class that that example belongs to. A classifier has a number of parameters that have to be learned from training data. The learned classifier is essentially a model of the relationship between the features and the class label in the training set. This relationship is tested by using the learned classifier on a different set of examples, the test data. The amount and quality of training data is definitely an important parameter affecting the performance of the classifier.

The research efforts and directions related to the present work were identified through literature survey. Rajeswari S and Theiva Jayaselvi K [1] have classified between normal and abnormal brain MR images using RBF kernel function of SVM classifier. The features for classification were extracted by Gray Level Co-occurance Matrix (GLCM) technique and the images were pre-processed by Median Filtering technique. The authors have achieved a classification accuracy of 65% by the RBF kernel and they conclude that, for large data, SVM may not work accurately due to training complexity. Mubashir Ahmed et. al. [2] developed a hybrid technique for the classification using Discrete Wavelet Transform (DWT), Principle Component Analysis (PCA) and SVM. PCA was used to select the best features for classification. These PCA selected features are given as an input to SVM for classification. Two SVM kernel functions – Linear kernel and Radial Basis kernel were used for classification. This approach has given a better result than earlier systems developed for the same purpose. Virendra Kumar Verma and Lalit P. Bhaiya [3] have designed a medical decision support using supervised neural network's Back Propagation Algorithm (BPA) for classification. They have experimented with three different sets of training and testing data taken from clump of images. Ahsan Bin Tufailet. al., [4] have automatically classified initial categories of Alzheimer's disease using SVM, k-NN and multilayer ANN to discriminate between the three classes. The performance of SVM, k-NN and ANN was found to be 60.65%, 68.06% and 53.57% respectively. They concluded that the KNN classifier is a good option for the overall classification and however, SVM classifier performed a lot better in identifying the true negatives (TN) than that of its counterparts.

The review of literatures has provided a good scope for the present study. New alternatives to keep the diagnosing methods much simpler and accurate are still being looked at. A critical review of literature has revealed that, SVM is less complex, more accurate, takes lesser computation time and have low generalization risk. Researchers had proposed various features for classifying tumor in MRI. According to Shantha Kumar and Ganesh Kumar [5], intensity feature alone is not sufficient; therefore other texture based features are to be extracted. Lerskiet. al. [6] in his research has indicated that most of the tumor is heterogeneous tissues and the mean values of relaxation times are not at all sufficient to characterize the heterogeneity of the different tumor types. An alternative approach, which can be investigated, is to apply texture analysis to the images to describe quantitatively the brightness and texture of the images. Review of literature indicates GLCM method is one of the important texture analysis techniques used for obtaining statistical properties for further classification.

Most of the researchers have used a few common and popular GLCM based Haralick texture features. An attempt till now has not been made to check the accuracy of classifiers by using all Haralick texture features. In reference 1, the authors declare that SVM may not work accurately for large training data due to training complexity. So, the amount of data in the training set is an important parameter for successful classification of SVM. Till now, investigations have not been made in the direction of identifying the optimal number of datasets to be used for training SVM. Hence, the objective of this is paper is to identify the correct number of training data for the SVM classifier for which a higher accuracy is obtained. The SVM classifier is input with 14 Haralick features extracted by GLCM.

METHODOLOGY

The architecture of the proposed system is illustrated in Figure 1. The methodology consists of five stages named asacquisition of brain MRI database, pre-processing of images, feature extraction, classification and performance analysis.

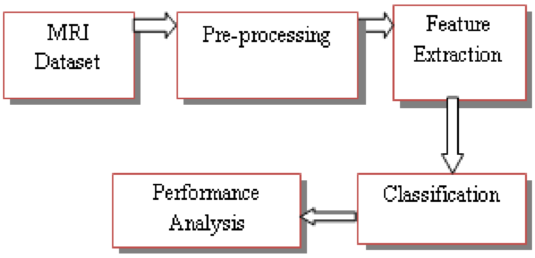

Figure 1: Proposed methodology

MRI DATASET

Datasets of Axial T2-weighted MR images is considered in our approach. A total of 64 patient's brain MRI images consisting of 10 normal, 32 malignant and 22 benign tumors were collected. Expert radiologist was consulted to confirm whether the tumor is malignant or benign.

PRE-PROCESSING OF IMAGES

Brain MR Images are subjected to be corrupted by noise during the image transmission and image digitization during the process of imaging. Pre-processing is a process to remove these noises from the MRI Brain image. The present work attempts to explore the use of median filtering technique to perform de-noising. In the median filtering operation, the pixel values in the neighbouring pixels are ranked according to intensity (brightness), and the median value (middle value) becomes the output value for the pixel under evaluation (central pixel

FEATURE EXTRACTION

Features are said to be properties that describes the whole image. There are varieties of features that can be extracted from images like Shape Features, Intensity features and Texture features. Texture features are mathematical parameters computed from the distribution of pixels, which characterize the texture type and thus the underlying structure of the objects shown in the image. The 14 texture features described and suggested by Haralicket. al. [7] include Angular second moment, Contrast, Correlation, Sum of squares, Inverse different moment, Sum average, Sum variance, Sum entropy, Entropy, Difference variance, Difference entropy, Information measures of correlation and Maximal correlation coefficient.

For texture based feature extraction GLCM technique is used in the current research work. GLCM contains information about the positions of pixels having similar gray level values. GLCM calculates the co-occurrence matrix of an image by computing how often a pixel with certain intensity 'i' occurs in relation to other pixel 'j' at a certain distance d and orientation. It is a two dimensional array, P, in which both the rows and the columns represent a set of possible image values.

SVM CLASSIFIER

In the case of SVM, a data point is viewed as a p-dimensional vector, and we want to know whether we can separate such points with a $(p-1)$-dimensional hyperplane. This is called a linear classifier. There are many hyperplanes that might separate the data. However, only one of them achieves maximum separation. One reasonable choice as the best hyperplane is the one that represents the largest separation, or margin, between the two classes. In addition to performing linear classification, SVMs can efficiently perform non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

RESULTS AND DISCUSSION

The objective of this paper is to investigate the effect of training data on the performance of SVM used for classification of brain MR images into three classes namely 'Normal', 'Benign' and 'Malignant'. The brain MR images data set were divided into training dataset and testing dataset. The amount of MR image data used to train the classifier is a vital parameter influencing its performance. More number of training than what is actually required may lead to overfitting problems in the classifier. There is no rule as such to decide on the training dataset for an optimal performance. Trail-and-error is the only method to determine the optimal amount of training data and hence various training dataset values will have to be experimented to determine the best one for classification before the classifier can be used in confidence for brain MR images classification. Hence, in this work various MR images classification experiments were carried out on SVM for 50%, 60%, 70%, 80% and 90% of total dataset as the training dataset. The remaining 50%, 40%, 30%, 20% and 10% of the dataset were used as test data to measure the classification accuracy.

Fourteen Haralick texture features were extracted from the complete MRI dataset including training and testing dataset. These features were used as the input to SVM based on which classification process is carried out. Three different kernel functions viz., linear kernel, polynomial kernel and RBF kernel were tried in this paper for classification. The SVM kernel functions were trained with various considered percentage of training data and then the trained model was tested with the testing dataset. The classification accuracy of these kernel functions for testing was measured.

Table 1: Performance of SVM kernels for different percentage of training data

Percentage of Training Data	Kernel Function	Classification Accuracy
50	Linear	50%
	Polynomial	90%
	RBF	80%
60	Linear	87.5%
	Polynomial	100%
	RBF	87.5%
70	Linear	88.88%
	Polynomial	100%
	RBF	94.44%
80	Linear	90%
	Polynomial	100%
	RBF	100%
90	Linear	83.33%
	Polynomial	100%
	RBF	83.33%

Table 1 gives the performance of SVM kernel functionsfor different percentage of training data. The performance of the classifier kernel functions is measured in terms of the classification accuracy. Figure 2, Figure 3 and Figure 4 shows the variation of classification accuracy with the percentage of training data for linear, polynomial and RBF kernel functions. Figure 5 is the plot showing the performance of all the three kernel functions in terms of classification accuracy.

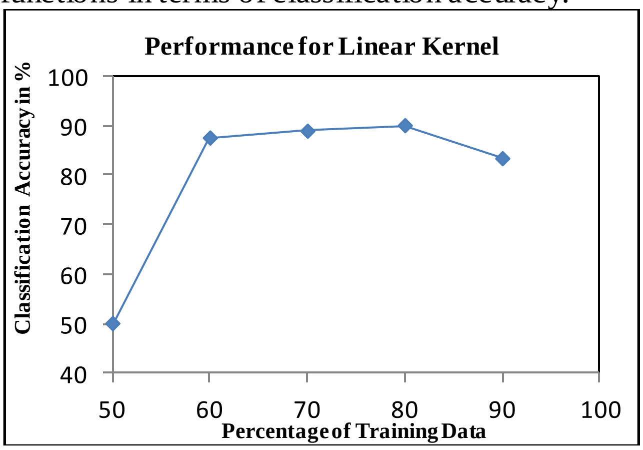

Figure 2: Performance of linear kernel for various percentages of training data

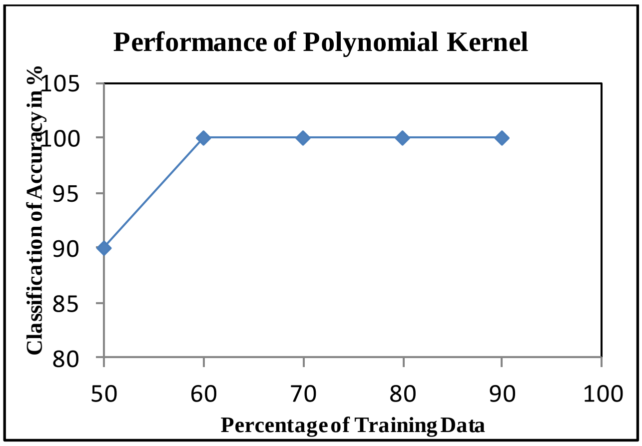

Figure 3: Performance of polynomial kernel for various percentages of training data

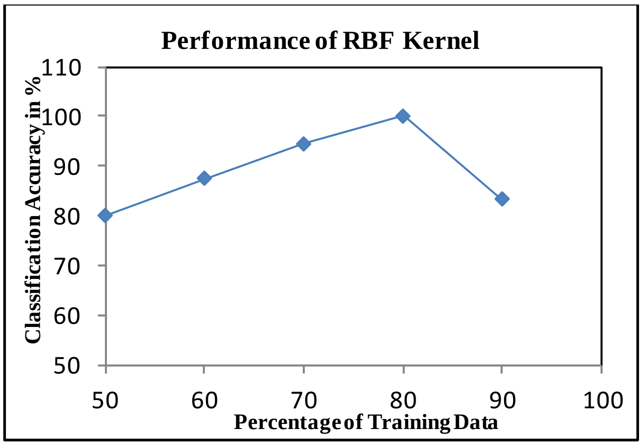

Figure 4: Performance of RBF kernel for various percentages of training data

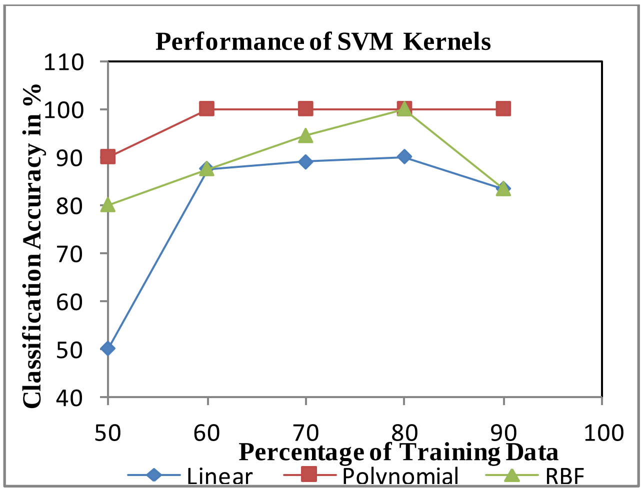

Figure 5: Performance of all kernels for various percentages of training data

Results indicate that the classification accuracy of linear kernel function varies from 50% to 90% for different percentage of training data. For polynomial kernel function the accuracy of classification is either 90% or 100% for various training data. Infact, for most of the training data the performance of polynomial kernel is 100%. The classification accuracy for RBF kernel function varies from 80% to 100%.

Figure 2 and Figure 4 indicates that the classification accuracy for linear and RBF kernel functions increases with the increase in the percentage of training data. The increasing trend in the accuracy can be observed upto 80% of training data, and further increase in the training data to 90% results in a sharp fall of the accuracy. Increase of the classification accuracy with the increase in the percentage of data in the training set is due to the fact that more number of data in the training set will make the classifiers to learn the process better and make better classifications. However, the drop in the accuracy at 90% of training data indicates that the classifier is overfit for the given training dataset. A model which has been overfit will generally have poor classification performance, as it can exaggerate minor fluctuations in the data. Overfitting occurs after 80% of data in the training set, where the classifier begins to memorize training data rather than learning to generalize from trend. Hence, the present study reveals that, for the considered dataset, 80% of training data is optimum, for which better classification accuracy can be obtained in linear and RBF kernel functions.

The classification accuracy of polynomial function reaches 100% for 60% of training data and the accuracy remains at 100% for further increase in the training data as shown in Figure 4. The overfitting problem of the classifier seems to be of less importance in polynomial kernel function.

Figure 5 shows the comparison of performances of the three types of SVM kernel functions used in the present work. It is always important that the research on classification of brain into different classes based on patients MRI images suggest one decisive method for this classification problem, so that the radiologists can be at ease. With this standpoint, when the results of the SVM classifier kernel functions are studied, polynomial kernel function is better than the other two as its classification accuracy is higher.

CONCLUSIONS

An automated intelligent classification system is proposed which caters the need for classification of image slices for tumor identification. To study the effect of the training data, the performance of linear, polynomial and RBF SVM kernel functions is examined for various percentages of training data. The best classification accuracy was obtained for 80% of data in the training set for all the three kernel functions. The polynomial kernel function of SVM was found to more accurately classify the brain MR images. Further studies in this field can focus on investigating the effect the training data on other widely used classifiers.

REFERENCES

1) Rajeswari S &TheivaJayaselvi K, Support Vector Machine Classification for MRI Images, International Journal of Electronics and Computer Science Engineering, 1(3), pp. 1534-1539, 2012.
2) Mubashir Ahmad, Mahmood ul-Hassan, Imran Shafi, Abdelrahman Osman, Classification of Tumors in Human Brain MRI using Wavelet and Support Vector Machine, IOSR Journal of Computer Engineering, 8(2), pp. 25-3, 2012,
3) Virendrakumarverma and Lalit P. Bhaiya, MRI Brain Images Classification using Soft Computing Techniques, International Journal of Computer Technology & Applications, 3 (3), pp.1179-1182.
4) Ahsan Bin Tufail, Ali Abidi, Adil Masood Siddiqui, and Muhammad ShahzadYounis, Automatic Classification of Initial Categories of Alzheimer's Disease from Structural MRI Phase Images: A Comparison of PSVM, KNN and ANN Methods, World Academy of Science, Engineering and Technology, 72, pp.1731-1735, 2012.
5) P. Shantha Kumar and P. Ganesh Kumar, Performance Analysis of Brain Tumor Diagnosis Based on Soft Computing Techniques, American Journal of Applied Sciences, 11(2), pp.329-336, 2014,.
6) R. A. Lerski, K. Straughan, L. R. Schad, D. Boyce, S. Bluml, and I. Zuna, MR Image Texture Analysis- An approach To Tissue Characterization, Magnetic Resonance Imaging, 11, pp.873-887, , 1993.
7) Robert M. Haralick, K. Shanmugam and It'shakDinstein, Texture Features for Image Classification, IEEE Transactions on Systems, Man and Cybernetics, 3(6), pp. 610-621, 1973.