Advanced Science and Technology Letters
Vol.98 (CES-CUBE 2015), pp.34-37
http://dx.doi.org/10.14257/astl.2015.98.09

Incremental Eigenspace Model Applied to Real World Problem

Byung Joo Kim

Youngsan University Department of Computer Engineering, Korea, bjkim@ysu.ac.kr

Abstract. Recent years have witnessed a dramatic increase in our ability to collect data from various sensors, devices, in different formats, from independent or connected applications. This data flood has outpaced our capability to process, analyze, store and understand these datasets. Many machine learning algorithms do not scale beyond data sets of a few million elements or cannot tolerate the statistical noise and gaps found in real-world data. Further research is required to develop algorithms that apply in real-world situations and on data sets of trillions of elements. In this paper we propose a scalable algorithm to handle the huge collections of data. Through the experimental results, proposed method performs well on huge data from UCI machine learning repository data Set.

Keywords: Incremental Eigenspace, Conjugate, Support Vector Machine

1 Introduction

Traditional machine learning has been largely concerned with developing techniques for small or modestly sized datasets. These techniques fail to scale up well for large data, a situation becoming increasingly common in today's world. Furthermore most of the machine learning classifiers are trained in a batch way. Under this model, all training data is given a priori and training is performed in one batch. If more training data is later obtained the classifier must be re-trained from scratch. Re-solving the problem from scratch seems computationally wasteful. In this research we will focus on developing classifier for big data sets and incremental way of learning for dealing with real world problem. In this paper we propose a new classifier for on-line and big data. Paper is composed of as follows. In Section 2 conjugate based LS-SVM method is described in Section 3. Experimental results to evaluate the performance of proposed classifier is shown in Section 4.

2 Conjugate LS-SVM for Real World Data

Support vector machines(SVM) developed by Vapnik [1] and it is a powerful methodology for solving problems in nonlinear classification[2]. Originally, it

ISSN: 2287-1233 ASTL

has been introduced within the context of statistical learning theory and structural risk minimization. In the methods one solves convex optimization problems, typically by quadratic programming(QP). Solving QP problem requires complicated computational effort and need more memory requirement. LS-SVM[3] overcomes this problem by solving a set of linear equations in the problem formulation. LS-SVM method is computationally attractive and easier to extend than SVM. But traditional batch way LS-SVM requires storing (N+1) × (N+1) matrix where N is a number of patterns. It is infeasible method when dealing with big data[4]. For big data sets the use of iterative methods is recommended. In principle, various methods can be used at this point including SOR(Successive Over-Relaxation), CG(Conjugate Gradient), GMRES(Generalized Minimal Residual) etc. However, not all of these iterative methods can be applied to any kind of linear system. For example, in order to apply CG the matrix should be positive definite. Due to the presence of the b bias term in the LS-SVM model the resulting matrix is not positive definite. So before we can apply such methods we have to transform the linear system into a positive definite system. The LS-SVM KKT system is of the form

$$\begin{bmatrix} 0 & Y^T \\ Y & H \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} \tag{1}$$

more specifically with $H = \Omega + I/\gamma$, $\xi_1 = b, \xi_2 = \alpha$, $d_1 = 0, d_2 = I_v$. This can be transformed into

$$\begin{bmatrix} s & 0 \\ 0 & H \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 + H^{-1} y \xi_1 \end{bmatrix} = \begin{bmatrix} -d_1 + y^T H^{-1} d_2 \\ d_2 \end{bmatrix} \tag{2}$$

with $S = y^T H^{-1} y > 0$ ($H = H^{-T} > 0$) Because s is positive and H positive definite the overall matrix is positive definite. This form is very suitable because different kinds of iterative methods can be applied to problems involving positive definite matrices. This leads to the LS-SVM classifier with conjugate gradient algorithm LS-SVM for big data is as follows.

1. Solve η, ν from $H\eta = Y$ and
   $H\nu = 1_v$
2. Compute $s = Y^T \eta$
3. Find solution
   $b = \eta^T 1_v / s$
   $\alpha = \nu - \eta b$

3 Experiment

To evaluate the performance of proposed classification system, experiment is performed on big data. First we evaluate the proposed system to HIV-1 protease cleavage data set.

3.1 Gas sensor array under dynamic gas mixtures Data Set

Gas sensor array under dynamic gas mixtures data set [5] contains 4178504 instances and number of attributes are 19. The first 80% are provided here as the training dataset and the remaining 20% as the testing dataset. In [6] it is shown that the use of 10-fold cross-validation for hyperparameter selection of LS-SVMs consistently leads to very good results.

Table 1. Training and generalization result on gas sensor array under dynamic gas mixtures data set

	Training	Generalization	Eigenvalue update criterion
Standard LS-SVM	100%	98.02%	none
Proposed method	100%	97.8%	$\lambda_i^{'} \rangle 0.7\overline{\lambda}$

The results on the Gas sensor array under dynamic gas mixtures data set are given in Table 1. Generalization ability in proposed method is similar to standard LS-SVM. But in standard LS-SVM (1138,563) x (1138,563) matrix is needed. It is not infeasible for big data.

4 Conclusion and Remarks

A conjugate based LS-SVM which combining incremental KPCA was presented for dealing with big data. Such classifier has following advantages. Proposed classifier is more efficient in memory requirement than batch LS-SVM. In batch LS-SVM the (N+1) × (N+1) matrix has to be stored, while for our proposed method does not. It is very useful when dealing with big data.Experimental results on huge data from UCI machine learning repository, proposed method shows lead to good performance.

References

1. Vapnik, V.N.: Statistical learning theory. John Wiley & Sons, New York (1998)

2. Winkeler, J., Manjunath, B.S., Chandrasekaran, S.: Subset selection for active object recognition. In CVPR, Vol. 2, IEEE Computer Society Press, June pp. 511--516 (1999)
3. Suykens, J.A.K., Vandewalle, J.: Least squares support vector machine classifiers. Neural Processing Letters, vol. 9, pp. 293--300 (1999)
4. Hall, P.D., Marshall, Martin, R.: Incremental eigenalysis for classification. In British Machine Vision Conference, Vol. 1, September pp. 286--295 (1998)
5. University of California Irvine Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/Gas+sensor+array+under+dynamic+gas+mixtures
6. Golub, G.H., Van, C.F.: Large scale LS_SVM Matrix Computations, Baltimore MD: Johns Hopkins University (2002)

