The Generalization Ability of SVM Classification Based on Markov Sampling

Meghna IIT2018109
Riya Chaudhary IIT2018145
Vidhi Sah IIT2018169
Nandini Goyal IIT2018173
Chaitali Agrawal IIT2018504

Semester VI, Department of IT, Indian Institute of Information Technology, Allahabad, India.

Abstract: In this paper we have proposed the use of markov sampling for improving the accuracy of SVM based classifiers, in the particular case of a dataset with a high number of features.

I. INTRODUCTION

An assumption have been made in the given algorithm that the data being fed to the SVC is Independent and Identically distributed. (I. I. D.). It is also known that this assumption is not always true specially when we are talking about data which is inherently temporal in nature, For example - market prediction, speech recognition, etc.

The algorithm uses a sampling method known as Markov Sampling which has been taken from Markov chain Monte Carlo (MCMC) methods. This method is used to replace the random sampling which is generally used to train the SVCs.

II. General Terms Used

1. SVM :

   SVM is a binary classification model developed by Vapnik from Structural Risk minimization theory. It uses the technique known as The Kernel trick in which it transforms the data and accordingly finds the optimal boundary between the possible outputs.

   Reasons for SVMs being so important are -

   - When a dataset is considered with large number of features and small samples size, SVMs are very powerful.
   - Using SVM, both simple and highly complex

classification models can also be learned.

- SVM is helpful in avoiding the overfitting of curve by utilizing advanced mathematical principles.

2. Markov sampling :

   In statistics, Markov chain Monte Carlo (MCMC) methods comprise a class of algorithms for sampling from a probability distribution. By constructing a Markov chain that has the desired distribution as its equilibrium distribution, one can obtain a sample of the desired distribution by recording states from the chain. The more steps are included, the more closely the distribution of the sample matches the actual desired distribution. Various algorithms exist for constructing chains, including the Metropolis–Hastings algorithm.

   Markov Sampling is inspired from Markov Chain Monte Carlo methods. In statistics it comprises of a class of algorithms for sampling from a probability distribution. If we construct a Markov chain with the desired distribution according to its equilibrium distribution, then we can obtain a sample of the desired distribution by recording states from the chain. The more steps are included, the more closely the distribution of the sample matches the actual desired distribution. Metropolis – Hastings algorithm is one of the various existing algorithms for construction of chains.

III. ALGORITHM DESCRIPTION

1. Let m be the size of training samples and m%2 be the remainder of m divided by 2. m^+ and m^-denote the size of training samples which label are +1 and −1, respectively. Draw randomly N1(N1 ≤ m) training samples {zi} N1 i=1 from the dataset Dtr. Then we can obtain a preliminary learning model f0 by SVMC and these samples. Set $m^+ = 0$ and $m^- = 0$.

2. Draw randomly a sample from Dtr and denote it the current sample zt. If m%2 = 0, set $m^+ = m^+ + 1$ if the label of zt is +1, or set $m^- = m^- + 1$ if the label of zt is −1.

3. Draw randomly another sample from Dtr and denote it the candidate sample z∗.

4. Calculate the ratio P of e−(f0,z) at the sample z∗ and the sample zt, P = e−(f0,z∗)/e−(f0,zt)

5. If P = 1, yt = −1 and y∗ = −1 accept z∗with probability P = e−y∗f0 /e−ytf0 . If P = 1, yt = 1 and y∗ = 1 accept z∗ with probability P = e−y∗f0 /e−ytf0 . If P = 1 and yty∗ = −1 or P < 1, accept z∗ with probability P. If there are k candidate samples z∗ can not be accepted continuously, then set P = qP and with probability P accept z∗. Set zt^{+1} = z∗, $m^+ = m^+ +1$ if the label of zt is +1, or set $m^- = m^- + 1$ if the label of zt is −1

[if the accepted probability P (or P, P) is larger than 1, accept z∗ with probability 1].

6. If $m^+ < m/2$ or $m- < m/2$ then return to Step 3, else stop it.

IV. RESULTS AND OBSERVATIONS

Kernel	Accuracy
Linear	78.43%
RBF	82.56%
X^2	70.96%
Optimal Hyperparameter (On RBF)	82.68%

V. CONCLUSION

In the step 2 of the algorithm it is taken into consideration about the number of samples being even or odd. It is done by making appropriate increments in the count of positive and negative classes.

In the proposed algorithm the constants are taken as k = 5 and q = 1.2 . To compute the transition probabilities P or P' or P", the model has to be initially trained by SVM classifier using a subset of training dataset.

The accuracy achieved using a non-linear kernel (approx 84 %) is much higher than that of a linear one (approx 79%). It can be concluded that the problem is highly non-linear in nature.

VI. REFERENCES

[1] Steinwart, "Consistency of support vector machines and other regularized kernel classifiers," IEEE Trans. Inf. Theory, vol. 51, no. 1, pp. 128–142, Jan. 2005.

[2] Steinwart and A. Christmann, "Fast learning from non-i.i.d. observations," in Proc. Adv. Neural Inf. Process. Syst., vol. 22. Vancouver, BC, Canada, Dec. 2009, pp. 1768–1776.

[3] T. Steinwart, D. Hush, and C. Scovel, "Learning from dependent observations," J. Multivariate Anal., vol. 100, no. 1, pp. 175–194, Jan. 2009.

[4] S. Smale and D. X. Zhou, "Online learning with Markov sampling," Anal. Appl., vol. 7, pp. 87–113, Jan. 2009.

[5] F. Cucker and S. Smale, "On the mathematical foundations of learning," Bull. Amer. Math. Soc., vol. 39, no. 4, pp. 1–49, Jan. 2001.