International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:4, 2009

A Kernel Classifier using Linearised Bregman Iteration

K. A. D. N. K Wimalawarne

Abstract—In this paper we introduce a novel kernel classifier based on a iterative shrinkage algorithm developed for compressive sensing. We have adopted Bregman iteration with soft and hard shrinkage functions and generalized hinge loss for solving l_1 norm minimization problem for classification. Our experimental results with face recognition and digit classification using SVM as the benchmark have shown that our method has a close error rate compared to SVM but do not perform better than SVM. We have found that the soft shrinkage method give more accuracy and in some situations more sparseness than hard shrinkage methods.

Keywords—Compressive sensing, Bregman iteration, Generalised hinge loss, sparse, kernels, shrinkage functions

I. Introduction

WITH recent research advances in Compressive Sensing [1], lot of attention towards research on solving l_1 norm equations of the type (1) have been received. Many areas such as signal processing, image processing and machine learning also find these types of problems highly important.

$$\min_{\mathrm{u}} |u|_1 + \|\mathrm{A}u - y\|_2^2 \tag{1}$$

In the recent past, a large number of approaches to solve these problems have been published and more new methods can be expected in the future. Among them Iterative Shrinkage Thresholding Algorithm (ISTA) [2], Fast Iterative Shrinkage Thresholding Algorithm (FISTA) [3], Bregman iteration [4] are few to mention. All these algorithms are based on soft shrinkage methods which has been found to be highly sucessful.

Many problems in machine learning use norm minimization for various reasons. Widely used and researched algorithms such as LASSO [5] and 1-norm SVM [6] have found it as a tool for sparseness and feature selections. In other recently developed areas such as self-taught learning [7] and sparse coding [8], the l_1 regularization has played an important role. It clearly shows that research on l_1 minimization problems has major importance to machine learning and development of compressive sensing can greatly influence and benefit research in many machine learning problems such as classification and regression.

Several machine learning researchers have already started to adapt novel compressive sensing algorithms for machine learning including Koh, Kim and Boyd [9] who have developed a Logistic regression based classifiers and Langford and Zhang [10] have used sparse gradient method for online learning. Inspired by these researches we have researched further developing a novel machine learning method using Bregman iterations method. Our focus was mainly on classification problem and we considered several aspects in designing a robust classifier. Due to its robustness we considered kernel classifier as SVM. Sparsity was another major factor we considered where much research was directed towards loss functions and shrinkage functions since they can be primarily important for achieving high sparsity.

The rest the paper is strutured in the following manner, the next section talks about the basic formulation of linearised Bragman iteration method, followed by loss funciton and shrinkage functions. Next section talks about experiments and results and the final section talks about conclusions and future works.

II. Linearied Bregman Iteration

Among many recently developed Compressive Sensing algorithms to solve (1) our attention was drawn towards Bregman iteration [4] mainly due to its simplicity and its flexibility in implementation. Closely related the Fixed Point Continuation theory [11] also has laid a solid foundations on iterative shrinkage mechanisms with global convergence. Bregman iteration has been discussed in several papers but we put forward the derivation here for the sake of completion and readers are suggested to refer [4] for a comprehensive treatment.

Considering the binary classification problem with data $\{x_i, y_i\}$ where $i = 1..n$ with x_i representing the data element and their labels $y_i \in \{-1, 1\}$ we can define an optimisation problem with 1-norm minimisation as follows, where w represent parameter to learn and L representing the loss function.

$$\min_{\mathrm{w}} |w|_1 + L(x, w, y) \tag{2}$$

In Compressive Sensing the loss function L is taken as $\|\mathrm{Aw}-y\|$ which is equivalent to the least square error where A contains the data elements. We can build a kernel classifier directly by replacing kernel elements with an appropriate kernel function $K(.,.)$ (Gaussian kernel) [12] as $A_{ij} = K(x_i, x_j)$. Though least square error type loss functions can be used for classification, the most commonly used is the hinge loss [13] which is a non-diferentiable function which makes it hard to devise a good solution. Alternatively a more suitable loss function that is differentiable in order to solve it easily such as the generalized hinge loss is discussed in Section 3 can be substituted.

Both machine learning and compressive sensing problems can be generally represented by equation(3).

K. A. D. N. K Wimalawarne is with the Department of Computer Science and Engineering, University of Moratuwa, Katubedda, Sri Lanka e-mail: (kishanwn@gmail.com).

$$\min_{\mathrm{u}} \{J(u) + H(u)\} \; J(u) = \mu \, ||u||_1 \; H(u) = \text{Loss Function} \tag{3}$$

Solving equation (3) can be a difficult task and as proposed in [4], linearisation is an efficient method in solving it. Using the Bregman distance [4] at step k of an iteration we can define a distance measure $D(u, u^k)$ as follows.

$$D(u, u^k) := J(u) - J(u^k) - < \partial J(u^k), u - u^k > \tag{4}$$

Using Taylor series we can also approximate the H as in (7) with an addition of a penaly term of $\frac{1}{2\delta}||u - u^k||^2$

$$\tilde{H}(u, u^k) = H(u^k) + < \nabla H(u), u - u^k > + \frac{1}{2\delta}||u - u^k||^2 \tag{5}$$

Combining both (4) and (5) one can design the value of u_{k+1} as a minimisation problem which can be further simplified to (6).

$$u^{k+1} = \mathrm{argmin}_{\mathrm{u}} D(u, u^k) + \tilde{H}(u, u^k) \tag{6}$$

Again by differentiating (6) with respect to u we arrive at (7) with $p^k = \partial J(u)$

$$0 = p^{k+1} - p^k + \nabla H(u^k) + \frac{1}{\delta}(u^{k+1} - u^k)) \tag{7}$$

By defining v as follows we can arrive at

$$v^k = p^k + \frac{1}{\delta} u^k \tag{8}$$

$$v^{k+1} = v^k - \delta \nabla H(u^k) \tag{9}$$

$$u^{k+1} = \delta \cdot \mathit{shrink}(v^{k+1}, \mu) \tag{10}$$

At each iteration u^{k+1} can be updated using a soft or a hard shinkage method as in equation (10) until convergence.

The training process can be put forward as a simple algorithms as given below.

Algorithm 1

```
$v^0 = 0$
while not converged
    $v^{k+1} = v^k - \delta \nabla H(u^k)$
    $u^{k+1} = \delta \cdot shrink(v^{k+1}, \mu)$
```

These type of shrinkage algorithms based on Bregman iterations have been explained and their converge convergence are proven in [4].

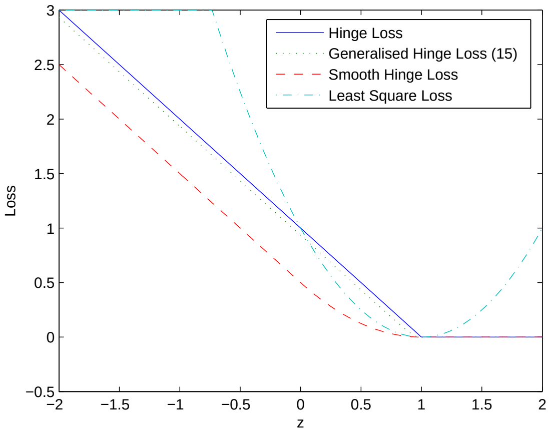

Fig. 1. Loss functions used for clasification problems

III. Loss Functions

Many factors such as accuracy and sparsity in machine learning problems depends on the selection of the loss function. Hinge loss [12] is commonly used with SVM for classification problems but since it is not differentiable it cannot be directly used in our approach. We experimented with several alternative loss functions such as the smooth hinge loss [13] and squared hinge loss [10] which have been used in place of the hinge loss but resulted in less accurate classifications and less sparse solutions.

In search of a better loss function we recognized the generalized hinge loss [13] as an ideal loss function (11) for our problem since it is differentiable as shown in equation (12) and can be well approximated to the hinge loss functions with higher values of α (Fig 1)

$$h_\alpha(z) = \begin{cases} \frac{\alpha-1}{\alpha} - z & \text{if} z \leq 0 \\ \frac{1}{\alpha+1} - z^{\alpha+1} + \frac{\alpha-1}{\alpha} & \text{if } 0 < z < 1 \\ 0 & \text{if } z \geq 1 \end{cases} \tag{11}$$

$$h'_\alpha(z) = \begin{cases} -1 & \text{if } z \leq 0 \\ z^\alpha - 1 & \text{if } 0 < z < 1 \\ 0 & \text{if } z \geq 1 \end{cases} \tag{12}$$

Fig. 1 shows the different loss functions and its clear that the generalized hinge loss approximate the hinge loss closely compared to other loss functions such as squared hinge and smooth hinge loss. This motivated us to use generalized hinge loss with higher ($\alpha = 15$) to build our classifier.

IV. Shrinkage Functions

In iterative shrinkage problems their shrinkage functions is a crucial component. Over the years researchers have proposed soft [4] and hard [5] shrinkage functions. For both Bregman iteration and Fixed Point methods soft shrinkage functions of the type (13) have been used and their global convergence has also been proven [4]. This type of shrinkage functions are commonly widely used throughout compressive sensing and inverse problems.

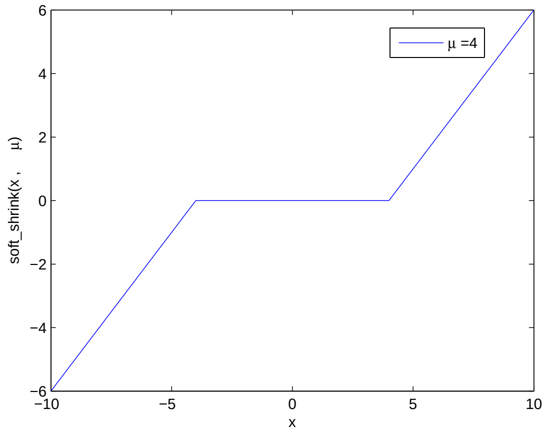

Fig. 2. Soft shrinkage with $\mu = 4$

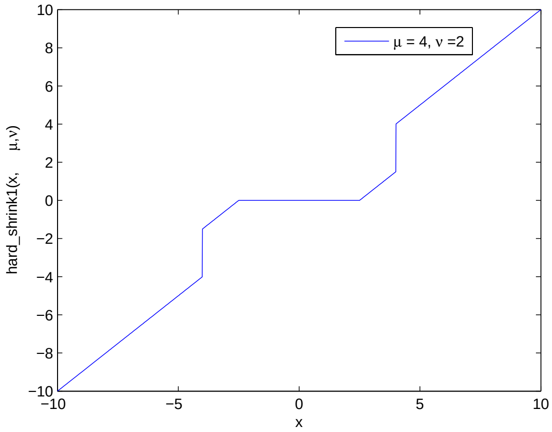

Fig. 3. Hard shrinkage with $\mu = 4$, $\nu = 2$

Inspired by [9] and [14] we experimented with alternative approaches of using hard shrinkage functions. Among hard shrinkage functions we found that truncation methods in [9] worked highly efficiently preserving the convergence of the problem and with higher sparsity. The Fig. 2 shows the soft shrinkage functions. Equaitons (14) and (15) show hard shrinkage methods and Fig. 3 and Fig. 4 show their respective graphs. μ and ν are user defined parameters to control thresholding. In the next section we present details of experiments that we carried with these shrinkage methods and thier results.

$$\begin{aligned} u_i^{k+1} &= \delta \cdot soft\_shrink(v_i^k, \mu) \\ &:= \delta \cdot \mathrm{sgn}(v_i^k) \max\{|v_i^k| - \mu\} \\ &= \begin{cases} v_i^k - \mu, & v_i^k \in (\mu, \infty) \\ 0, & v_i^k \in [-\mu, \mu] \\ v_i^k + \mu, & v_i^k \in (-\infty, -\mu) \end{cases} \end{aligned} \tag{13}$$

$$u_i^{k+1} = \delta \cdot hard\_shrink1(v_i^k, \mu, \nu)$$

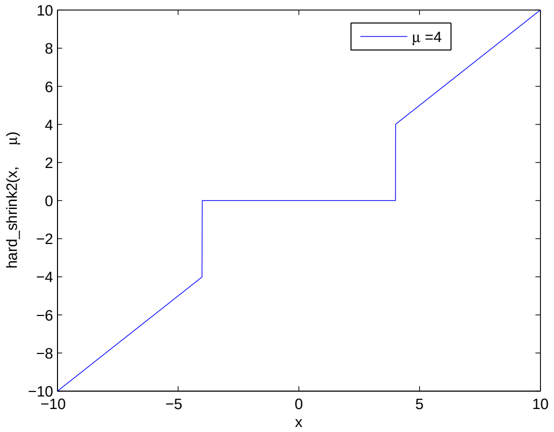

Fig. 4. Hard shrinkage with $\mu = 4$

$$\begin{aligned} &:= \delta \cdot \mathrm{sgn}(v_i^k) \max\{|v_i^k| - \mu\} \\ &= \begin{cases} v_i^k - \mu, & v_i^k \in (\nu, \infty) \\ 0, & v_i^k \in [-\nu, \nu] \\ v_i^k + \mu, & v_i^k \in (-\infty, -\mu) \end{cases} \end{aligned} \tag{14}$$

$$u_i^{k+1} = \delta \cdot hard\_shrink2(v_i, \mu) = \begin{cases} 0 & v_i \in [-\mu, \mu] \\ v_i & otherwise \end{cases} \tag{15}$$

V. Experiments

We experimented our proposed classifiers with face recognition using Shefield face database [15]. Our data set was relatively small since it only used 100 training images and 575 test images. Our proposed classifier was used Gaussian kernels [12] with data and for comparisons we used the standard SVM Adatron classifiers [16]. Table 1 shows the results obtained from experiments where it shows the classification accuracies and average sparsity with different shrinkage functions of the new method as well as the SVM. We used $\delta = 0.04$ and experiment with different values of μ and ν.

TABLE I
Classification Results of Faces

Classifier	Avg. Sparsity	Error %
Bregman with $soft\_shrink$	48	0.93
Bregman with $hard\_shrink1$	59	1.12
Bregman with $hard\_shrink2$	28	1.20
SVM	38	0.60

As we can see from the Table 1 our proposed classifier perform well as a classifier since they all have less error rates around 1%. But none of them performed better than the standard SVM. Interestingly the $soft\_shrink$ method has shown less error compared to both hard shrinkage methods. Among the hard shrinkage methods $hard\_shrink1$ has given less a error rate compared to the other. When comparing sparsity the $hard\_shrink2$ method has given higher sparsity though it gives a worse error rate.

As our second experiment we used NMIST hand written digit classificaiton [17] using our method and SVM. As our training set we took 100 images from digits and trained different classifiers. For testing we took another 100 images to check the classification accuracy of each classifier as shown in Table 2. Again we see that none of our algorithms can outperform SVM though in general they all have close accuracy compared to SVM. As in the easier case the $soft\_shrink$ functions is better in accuracy than the other two methods and in this case it also achives a higher sparsity than other methods as well.

TABLE II
CLASSIFICATION RESULTS OF DIGITS

Classifier	Avg. Sparsity	Error %
Bregman with $soft\_shrink$	248	2.24
Bregman with $hard\_shrink1$	342	2.36
Bregman with $hard\_shrink2$	285	2.38
SVM	241	1.77

VI. CONCLUSIONS AND FUTURE WORK

In this paper we have proposed a novel kernel based classifier that based on Bregman iteration. Our experiments have shown that our approach is a good classifier with high accuracy but it cannot outperform the standard SVM classifiers. Novelties that we propose in using generalized hinge losses and hard thresholding has proven to achieve highly sparse results. Further research can be carried out in designing shrinkage methods probably with adaptive shrinkage limits that can benefit in generating even more sparse results and more accuracy.

One major aspect that we need to be investigated in connection with our methods is its suitability with large scale problems. In the recent published research in compressive sensing and optimization there has been many methods proposed to solve one norm minimization problems with faster convergence rates. Among them Nestorov gradient [3] methods are highly promising methods.We believe that further research in adapting these methods to build hybrid algorithms in connection to ours would result in more efficient classifiers that can be efficiently used for practical applications specially to be used with large scale problems.

REFERENCES

[1] Donoho. D.L.: “Compressed sensing”, *IEEE Trans. Inform. Theory*, 52:1289-1306, (2006)

[2] Daubechies I., Defrise M., De Mol C. “An iterative thresholding algorithm for linear inverse problems with a sparsity constraint”, *Comm. Pure Appl. Math.* 57, pp. 14131457 (2004)

[3] Beck A., Teboulle M.,“A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems”, *SIAM J. Imaging Sciences*, forthcoming.

[4] Yin W., Osher S., Goldfarb D., Darbon J., “Bregman Iterative Algorithms for l_1-Minimization with Applications to Compressed Sensing”, *SIAM J. IMAGING SCIENCES* Vol. 1, No. 1, pp. 143168 (2008)

[5] Tibshirani R., “Regression selection and shrinkage via the lasso”,*J. R. Stat. Soc. Ser. B 58*, pp. 267288 (1996)

[6] Zhu J., Rosset S., Hastie T., Tibshirani R., “1-norm Support Vector Machines”, *Advances in Neural Information Processing Systems 16*, (2003)

[7] Raina R., Battle A., Lee H., Packer B., Ng. A. Y., “Self-taught learning: Transfer learning from unlabeled data”, *ICML 2007*

[8] Lee H., Battle A., Raina R., Ng. A. Y., “Efficient sparse coding algorithms”, In *Advances in Neural Information Processing Systems 19 (NIPS-06)*, pages 801808, (2007)

[9] Langford J., Li L. Zhang T., “Sparse Online Learning via Truncated Gradient”, *Journal of Machine Learning Research* Vol. 10, pp. 777-801 (2009)

[10] Koh K., Kim S., Boyd S., “An Interior-Point Method for Large-Scale l1-Regularized Logistic Regression”, *Journal of Machine Learning Research* Vol. 8, pp. 1519-1555, 2007.

[11] Hale E., Yin W., Zhang. Y., “A Fixed-point continuation method for l_1-regularization with application to compressed sensing”, CAAM Technical Report TR07-07, Rice University, Houston, TX, 2007.

[12] Vapnik V. N., *Statsitical Leanring Theory*, Wiley Interscience, (1998)

[13] Rennie J.D.M., “Smooth Hinge Classfication”, Technical report, MIT (2005)

[14] Bredies K., Lorenz D. A., “Iterated hard shrinkage for minimization problems with sparsity constraints”, *SIAM Journal on Scientific Computing*,Vol. 30(2), pp. 657-683, 2008.

[15] Graham D. B., Allinson N. M., “Characterizing Virtual Eigensignatures for General Purpose Face Recognition”, in *Face Recognition: From Theory to Applications , NATO ASI Series F, Computer and Systems Sciences, Vol. 163. H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman-Soulie and T. S. Huang (eds)*, pp 446-456, 1998.

[16] Campbell C., Cristianini N., “Simple Learning Algorithms for Traning Support Vector Machines”, Technical Report, UNiversity of Bristol, 1998.

[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. “Gradient-based learning applied to document recognition”, *Proceedings of the IEEE*, 86(11):2278-2324, November 1998.