Unsupervised and Semisupervised Classification via Absolute Value Inequalities

Glenn M. Fung [*] & Olvi L. Mangasarian [†]

Abstract

We consider the problem of classifying completely or partially unlabeled data by using inequalities that contain absolute values of the data. This allows each data point to belong to either one of two classes by entering the inequality with a plus or minus value. By using such absolute value inequalities (AVIs) in linear and nonlinear support vector machines, unlabeled or partially labeled data can be successfully partitioned into two classes that capture most of the correct labels dropped from the unlabeled data.

Keywords: unsupervised classification, absolute value inequalities, support vector machines

1 Introduction

We begin by giving the following simple example of our unsupervised classifier based on the following linear classifying plane:

$$x'w - \gamma = 0, \tag{1.1}$$

where the column vector x represents any data point in an n-dimensional space R^n, $w \in R^n$ is the normal vector to the classifying plane, γ determines the distance from the origin of the plane, and the prime denotes the transpose of the column vector x. The plane (1.1) can be utilized to divide R^n into two disjoint halfspaces by the following two linear inequalities:

$$\begin{array}{rcl} x'w - \gamma & \geq & 1, \\ x'w - \gamma & \leq & -1. \end{array} \tag{1.2}$$

The key to our approach is to represent the last two inequalities by the following single absolute value inequality (AVI):

$$|x'w - \gamma| \geq 1, \tag{1.3}$$

where $|\cdot|$ denotes the absolute value. Thus if $x'w - \gamma \geq 0$ then AVI (1.3) reduces to the first linear inequality of (1.2), whereas if $x'w - \gamma \leq 0$ then AVI (1.3) reduces to the second linear inequality of (1.2). Thus if we impose the AVI (1.3) on an unlabeled dataset, the dataset will be divided into two categories to best fit the AVI (1.3) or equivalently the two linear inequalities (1.2).

There have been approaches to semisupervised classification utilizing support vector machines such as [4, 8], but none of them have utilized absolute value inequalities. As far as unsupervised classification is concerned there have been no approaches such as ours that do not use any labeled data as we do here. There are of course many clustering techniques such as [1, 3, 5, 6, 18] that do not utilize any labels. However these approaches are quite different from what we are proposing here. Furthermore none of

[*] Business & Customer Operations Unit, American Family Insurance, Madison, WI 53783. *gfung@amfam.com.*

[†] Computer Sciences Department, University of Wisconsin, Madison, WI 53706 and Department of Mathematics, University of California at San Diego, La Jolla, CA 92093. *olvi@cs.wisc.edu.*

these approaches utilize either absolute value inequalities [14] or absolute value equations [20, 21, 17, 12]. Our approach here, as will be made clear in Section 2, is based on dual complementarity, a technique utilized in [15] for the solution of the NP-hard absolute value equation by a few linear programs, typically three or less. More recently a different approach to the one described here, both in theory and method of solution [16], was proposed that utilizes absolute value inequalities to handle both labeled and unlabeled data. Similarly, the present approach can handle both such data, but computationally we shall restrict ourselves to the more difficult and interesting case of unlabeled data classification.

We briefly describe now the contents of the paper. In Section 2 we outline the theory behind our approach and in Section 3 we state our iterative algorithm that consists of solving a succession of linear programs with modified objective functions. In Section 4 we give computational results that show the effectiveness of our approach by recovering most labels that have been dropped from all (unsupervised) data points. Section 5 concludes the paper.

We describe now our notation. The scalar product of two column vectors x and y in a n-dimensional real space will be denoted by $x'y$. For a vector u, u_j represents the jth group of components, whereas u^i represents the ith iterate of u during an iterative process. The identity matrix in a real space of arbitrary dimension will be denoted by I, while a column vector of ones of arbitrary dimension will be denoted by e and a column of zeros by 0. The abbreviation "s.t." stands for "subject to".

2 Unsupervised and Semisupervised Classification

We begin with an unlabeled dataset consisting of m points in the n-dimensional space R^n represented by the $m \times n$ matrix A and the labeled dataset consisting of k points in R^n represented by the $k \times n$ matrix H and labeled by the $k \times k$ diagonal matrix D with entries of ± 1 which denote which class of $+1$ or -1 each row of H belongs to. Thus we wish to find two planes $x'w - \gamma = \pm 1$ in R^n that specify the ± 1 feasible regions generated by the two inequalities of (1.2) and which satisfy with minimal error vectors y and s the following inequalities:

$$\begin{array}{rcl} |Aw - e\gamma| + y & \geq & e, \\ D(Hw - e\gamma) + s & \geq & e, \\ (y, s) & \geq & 0, \end{array} \tag{2.4}$$

where the nonnegative slack variables (y, s) are to be driven to zero by the following optimization problem:

$$\begin{array}{rccl} & \displaystyle\min_{w,\gamma,y,s,z} e'z + \nu(e'y + e's) & & \\ \text{s.t.} & |Aw - e\gamma| + y & \geq & e, \\ & D(Hw - e\gamma) + s & \geq & e, \\ -z \leq & w & \leq & z, \\ & (y, s) & \geq & 0. \end{array} \tag{2.5}$$

Here in addition, the 1-norm of w is minimized in order to maximize the distance between the two bounding planes $x'w - \gamma = \pm 1$ of the ± 1 feasible regions of the inequalities of (1.2), while ν is a positive parameter that balances the two groups of objectives of (2.5). Without the variable y and the first absolute value inequality of the above minimization problem, (2.5) is a standard linear support vector machine classification problem [10, 11]. In order to treat the nonlinear absolute value inequality in (2.5) above we proceed as follows. We replace the term $|Aw - e\gamma|$ in the absolute value inequality by an upper bound r on it: $-r \leq (Aw - e\gamma) \leq r$ which results in the following linear programming problem:

$$\begin{array}{rccl}
 & \min\limits_{w,\gamma,y,s,z,r} h'r + \mu e^{'}z + \nu(e^{'}y + e^{'}s) & & \\
\text{s.t.} - r \leq & Aw - e\gamma & \leq & r, \\
 & r + y & \geq & e \\
 & D(Hw - e\gamma) + s & \geq & e, \\
-z \leq & w & \leq & z, \\
 & (y, s) & \geq & 0,
\end{array} \tag{2.6}$$

where μ is another positive parameter, and the vector h is determined by using dual complementarity as in [15], which will iteratively generate a solution to the original problem (2.5) as we shall describe in the next section.

3 Iterative Linear Programming Solution of Unsupervised & Semisupervised Classification

We begin by rewriting the last optimization problem (2.6) as follows:

$$\begin{array}{lccccccccl}
\min\limits_{w,\gamma,y,s,z,r} & 0'w & +0\gamma & +\mu e'z & +\nu e's & +\nu e'y & +h'r & & \\
\text{s.t.} & Aw & -e\gamma & & & & +r & \geq & 0, \\
 & -Aw & +e\gamma & & & & +r & \geq & 0, \\
 & & & & & y & +r & \geq & e, \\
 & DHw & -De\gamma & & +s & & & \geq & e, \\
 & w & & +z & & & & \geq & 0, \\
 & -w & & +z & & & & \geq & 0, \\
 & & & & s & & & \geq & 0, \\
 & & & & & y & & \geq & 0.
\end{array} \tag{3.7}$$

The dual to the above linear program (3.7) is the following linear program with nonnegative dual variables $(u_1, \ldots, u_6)$ corresponding to the first six inequalities of (3.7):

$$\begin{array}{lccc}
\max\limits_{u_1,u_2,u_3,u_4,u_5,u_6} & e'u_3 + e'u_4 & & \\
\text{s.t.} & A'u_1 - A'u_2 + H'Du_4 + u_5 - u_6 & = & 0, \\
 & -e'u_1 + e'u_2 - e'Du_4 & = & 0, \\
 & u_5 + u_6 & = & \mu e, \\
 & u_4 & \leq & \nu e, \\
 & u_3 & \leq & \nu e, \\
 & u_1 + u_2 + u_3 & = & h, \\
 & (u_1, u_2, u_3, u_4, u_5, u_6) & \geq & 0.
\end{array} \tag{3.8}$$

From the dual complementarity condition we have that:

$$(u_1)'(Aw - e\gamma + r) + (u_2)'(-Aw + e\gamma + r) = 0. \tag{3.9}$$

It follows from the nonnegativity of all bracketed terms in (3.9) that:

$$u_1 + u_2 > 0 \implies r = -Aw + e\gamma \text{ or } r = Aw - e\gamma \iff r = |Aw - e\gamma|. \tag{3.10}$$

Hence it follows from the last equality constraint of the dual linear program (3.8) and (3.10) above that:

$$h - u_3 > 0 \iff u_1 + u_2 > 0 \implies r = |Aw - e\gamma|. \tag{3.11}$$

It follows from the last implications that:

$$h = u_3 + \epsilon e \implies r = |Aw - e\gamma|, \tag{3.12}$$

where ϵ is some small positive parameter. This fact implies that if $h = u_3 + \epsilon e$ then the objective term $h'r + \nu e'y$ in the linear program (2.6) has minimized the term $h'|Aw - e\gamma| + \nu e'y$ which in effect is equivalent to minimizing the term $\nu e'y$ in the optimization problem (2.5). However we cannot *a priori* enforce the condition $h = u_3 + \epsilon e$ since we do not know what u_3 before solving the linear program (3.7). But, this gives us a lead to an iterative process, similar to that of [15], where h is reset to $h = u_3 + \epsilon$ once u_3 is computed with a previous value of h. This leads to the following iterative algorithm.

Algorithm 3.1. *Choose a parameter value ϵ for the definition of h in (3.12) (typically $\epsilon = 10^{-6}$), and a maximum number of iterations itmax (typically itmax$= 10$).*

(I) Initialize the algorithm by choosing an initial nonnegative random vector in R^m for h and solve the linear program (2.6). Set iteration number $i = 0$.

(II) If $r^i = |Aw^i - e\gamma^i|$ or i=itmax stop.

(III) Obtain a dual optimal variable $u^{i+1} \in R^{3m+k+2n}$, by solving the linear program (3.7) with $h = u_3^i + \epsilon e$.

(IV) Set $i = i + 1$ and go to Step (II).

We note that if at any iteration i in the above algorithm, if $u_3^{i+1} = u_3^i$, then $r^i = |Aw^i - e\gamma^i|$ and the plane $x'w^i - \gamma^i = 0$ divides R^n into two halfspaces each containing part of the unlabeled data represented by the matrix A within an error margin y^i, and the correct part of the labeled data represented by the matrix H within an error margin denoted by s^i. We thus have the following proposition.

Proposition 3.1. *If at some iteration i of Algorithm 3.1, $u_3^{i+1} = u_3^i$, then $r^i = |Aw^i - e\gamma^i|$ and the plane $x'w^i - \gamma^i = 0$ divides R^n into two halfspaces each containing part of the unlabeled data represented by the matrix A and the correct part of the labeled data represented by the matrix H within error margins denoted by y^i and s^i respectively.*

We note that all the above results can be extended to nonlinear kernel classification [19, 9, 7] by replacing the linear separating plane (1.1) by a nonlinear separating surface $K(x', A')u - \gamma = 0$ that is linear in the unknowns (u, γ), but nonlinear in the data variable $x \in R^n$, where K is any nonlinear kernel.

4 Computational Results

We first note that our linear program (2.6) can be utilized in a semi-supervised algorithm wherein both labeled and unlabeled data are present. However our experimental results will focus on the unsupervised classification case where no labels are available. This is mainly because of our novel absolute value inequality technique and the availability of various approaches for handling semisupervised classification. We shall implement our algorithm on several publicly available datasets from the UCI repository [24]: the Wisconsin Diagnosis Breast Cancer (WDBC), Ionosphere and Cleveland, as well as the federalist papers dataset as described in [2].

Since our main goal is unsupervised binary classification, we intentionally ignored the labels for each one of these datasets. However, the labels were utilized to measure the classification performance of our Algorithm 3.1. All the experiments were performed using MATLAB [22], on a Dell Latitude E5430 with a Intel Core i5-3340M Processor and 8 GB of RAM memory.

In order to have an idea of the effectiveness of the approach, we used K-means clustering [23] (the MATLAB-included implementation) as a baseline approach. Note that since our formulation (2.6) uses the L_1 norm for regularization of the w vector of the separating hyperplane, hence our algorithm produces a sparse solution. That is the generated classifier depends on a minimal subset, as little as 9%, of the original input feature components of the dataset. In contrast, K-means utilizes the entire components of the dataset.

Our algorithm requires tuning of the parameters ν and μ, however in an unsupervised formulation there are no labels to test the "goodness" of the parameters in the tuning process. To circumvent this difficulty, we use the Silhouette method, a widely used method of interpretation and validation of clusters of data [13]. The Silhouette coefficient is a number between zero and one that can be interpreted as a measure of how appropriately the data has been grouped or clustered. We favored combinations of the pair (ν^*,μ^*) that maximize the Silhouette coefficient. After we picked the optimal pair (ν^*,μ^*) we used the previously ignored labels to assess performance. We call our method Unsupervised Absolute Value Inequality Classifier **UAVIC**.

Table 1 shows the dimensions of each dataset, the performance **UAVIC** for the best (ν^*,μ^*) (best tuned performance) as described above. Since **UAVIC** performs unsupervised feature selection we report the number of used features in parentheses next to the corresponding performance. For comparison purposes the best performance achieved by any (ν,μ) pair and the K-Means performance (using all features) are also included.

Dataset	**(m , n)**	**UAVIC-tuned accuracy (# of features)**	**UAVIC-best accuracy (# of features)**	**K-means accuracy**
WDBC	(603,9)	0.96 (7)	0.96(7)	0.96
Ionosphere	(351,33)	0.69 (12)	0.73 (3)	0.70
Cleveland	(297,13)	0.73(5)	0.73(5)	0.84
Federalist	(106,70)	0.75(13)	0.75(13)	0.82

Table 1: Computational results for 4 datasets showing dimensions of each dataset, the performance of **UAVIC** for the best (ν^*,μ^*) (best tuned performance) , the best performance achieved by any (ν,μ) and the K-Means performance. Performance given in percentage correctness is based on classification correctness using original dataset labels not available to the unsupervised classification Algorithm 3.1.

5 Conclusion and Outlook

We have proposed an absolute-value-inequality-based classification of a totally or partially unlabeled dataset and tested it on four datasets without making use of any of the dataset labels in the algorithm. The proposed approach classification correctness based on data labels withheld from the algorithm ranged between 73% and 96%. Feature reduction achieved was in the range of 9% to 77%. Future research into other approaches for unlabeled data classification that utilize absolute value inequalities will hopefully lead to algorithms with higher classification accuracy.

References

[1] K. Al-Sultan. A Tabu search approach to the clustering problem. *Pattern Recognition*, 28(9):1443–1451, 1995.

[2] Glenn Fung. The Disputed Federalist Papers: SVM Feature Selection via Concave Minimization. Proceedings of the 2003 Conference on Diversity in Computing, 2003. http://dl.acm.org/citation.cfm?id=948551.

[3] M. R. Anderberg. *Cluster Analysis for Applications*. Academic Press, New York, 1973.

[4] K. P. Bennett and A. Demiriz. Semi-supervised support vector machines. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems -10-*, pages 368–374, Cambridge, MA, 1998. MIT Press.

[5] P. S. Bradley, O. L. Mangasarian, and W. N. Street. Clustering via concave minimization. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems -9-*, pages 368–374, Cambridge, MA, 1997. MIT Press. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/96-03.ps.

[6] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28:781–793, 1995.

[7] G. Fung and O. L. Mangasarian. Proximal support vector machine classifiers. In F. Provost and R. Srikant, editors, *Proceedings KDD-2001: Knowledge Discovery and Data Mining, August 26-29, 2001, San Francisco, CA*, pages 77–86, New York, 2001. Association for Computing Machinery. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/01-02.ps.

[8] G. Fung and O. L. Mangasarian. Semi-supervised support vector machines for unlabeled data classification. *Optimization Methods and Software*, 15:29–44, 2001. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/99-05.ps.

[9] G. Fung, O. L. Mangasarian, and A. Smola. Minimal kernel classifiers. *Journal of Machine Learning Research*, pages 303–321, 2002. University of Wisconsin Data Mining Institute Technical Report 00-08, November 200, ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/00-08.ps.

[10] O. L. Mangasarian. Generalized support vector machines. In A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 135–146, Cambridge, MA, 2000. MIT Press. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-14.ps.

[11] O. L. Mangasarian. Data mining via support vector machines. In E. W. Sachs and R. Tichatschke, editors, *System Modeling and Optimization XX*, pages 91–112, Boston, MA, 2003. Kluwer Academic Publishers. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/01-05.ps.

[12] O. L. Mangasarian. Absolute value equation solution via concave minimization. *Optimization Letters*, 1(1):3–8, 2007. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/06-02.pdf.

[13] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, Vol. 20, No. 1. (November 1987), pp. 53-65, doi:10.1016/0377-0427(87)90125-7.

[14] O. L. Mangasarian. Absolute value programming. *Computational Optimization and Applications*, 36(1):43–53, 2007. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/05-04.ps.

[15] O. L. Mangasarian. Absolute value equation solution via dual complementarity. Technical Report 11-03, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, September 2011. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/11-03.pdf. Optimization Letters 7(4), 2013, 625-630.

[16] O. L. Mangasarian. Unsupervised classification via convex absolute value inequalities. Technical Report 14-01,Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, March 2014. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/14-01.pdf.

[17] O. L. Mangasarian and R. R. Meyer. Absolute value equations. *Linear Algebra and Its Applications*, 419:359–367, 2006. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/05-06.pdf.

[18] O. L. Mangasarian and E. W. Wild. Feature selection in k-median clustering. Technical Report 04-01, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, January 2004. SIAM International Conference on Data Mining, Workshop on Clustering High Dimensional Data and Its Applications, April 24, 2004, La Buena Vista, FL, Proceedings Pages 23-28. http://www.siam.org/meetings/sdm04. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/04-01.pdf.

[19] O. L. Mangasarian and E. W. Wild. Nonlinear knowledge-based classification. *IEEE Transactions on Neural Networks*, 19:1826–1832, 2008. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/06-04.pdf.

[20] J. Rohn. Systems of linear interval equations. *Linear Algebra and Its Applications*, 126:39–78, 1989. http://www.cs.cas.cz/ rohn/publist/47.doc.

[21] J. Rohn. On unique solvability of the absolute value equation. *Optimization Letters*, 3:603–606, 2009.

[22] MATLAB,The MathWorks, Inc. A MATLAB Primer. http://www.mathworks.com/help/pdf_doc/matlab/getstart.pdf.

[23] Spath, H. Cluster Dissection and Analysis: Theory, FORTRAN Programs, Examples. Translated by J. Goldschmidt. New York: Halsted Press, 1985.

[24] UCI Machine Learning Repository University of California, Irvine. http://archive.ics.uci.edu/ml/.