Coulomb Classifiers:
Reinterpreting SVMs as Electrostatic Systems

Sepp Hochreiter and Michael C. Mozer
Technical Report CU-CS-921-01
Department of Computer Science
University of Colorado
Boulder, CO 80309–0430
{hochreit,mozer}@cs.colorado.edu

May 2001

Abstract

We introduce a family of classifiers based on a physical analogy to an electrostatic system of charged conductors. The family, called *Coulomb classifiers*, includes the two best-known support-vector machines (SVMs), the ν–SVM and the C–SVM. In the electrostatics analogy, a training example corresponds to a charged conductor at a given location in space, the classification function corresponds to the electrostatic potential function, and the training objective function corresponds to the Coulomb energy. The electrostatic framework not only provides a novel interpretation of existing algorithms and their interrelationships, but it suggests a variety of new methods for SVMs including kernels that bridge the gap between polynomial and radial-basis functions, objective functions that do not require positive-definite kernels, regularization techniques that are not cast in terms of violation of margin constraints, and speed-up techniques using either approximate or restricted-but-exact algorithms. Based on the framework, we propose novel SVMs and perform simulation studies to show that they are comparable or superior to standard SVMs. The electrostatic framework subsumes not only SVMs but also nearest neighbor, density estimation, vector quantization, and clustering techniques.

1 Introduction

Recently, Support Vector Machines (SVMs) [1, 8, 5] have attracted much interest in the machine-learning community and are considered state of the art for classification and regression problems. One appealing property of SVMs is that they are based on a convex optimization problem, which means that a single minimum exists and can be computed efficiently. In this paper, we present a new derivation of SVMs by analogy to an electrostatic system of charged conductors. The electrostatic framework not only provides a physical interpretation of SVMs, but it also gives insight as to some of the seemingly arbitrary aspects of SVMs (e.g., the diagonal elements in the quadratic form), and it allows us to derive novel SVM approaches.

We will discuss the classification of an *input vector* $x \in \mathcal{X}$ into one of two categories, "+" or "−". We assume a supervised learning problem in which N training examples are available, each example i consisting of an input x_i and a label $y_i \in \{-1, +1\}$.

We will introduce three electrostatic models that have direct analogy to machine-learning (ML) classifiers, starting with a relatively limited electrostatic model and the following two building on and generalizing from the previous. For each model, we describe the physical system and show its correspondence to an ML classifier.

1.1 Electrostatic model 1: Uncoupled point charges

Consider an electrostatic system of point charges populating a space $\mathcal{X}'$ homologous to $\mathcal{X}$. Each point charge corresponds to a particular training example; point charge i is fixed at location x_i in $\mathcal{X}'$, and has a charge of sign y_i. We define two sets of fixed charges: $S^+ = \{x_i \mid y_i = +1\}$ and $S^- = \{x_i \mid y_i = -1\}$. The charge of point i is denoted $Q_i \equiv y_i \, \alpha_i$, where $\alpha_i \geq 0$ is the amount of charge, to be discussed below.

We briefly review some elementary physics. If a unit positive charge is at x in $\mathcal{X}'$, it will be attracted to all charges in S^+ and repelled by all charges in S^-. To move the charge from x to $\tilde{x}$, the force must be overcome at every point along the trajectory; the path integral of the force along the trajectory is called the *work* and does not depend on the trajectory. The *potential* at x is the work that must be done to move a unit positive charge from a reference point (usually infinity) to x.

The potential at x is $\varphi(x) = \sum_{i=1}^{N} Q_i \, G(x_i, x)$, where G is a kernel measuring the distance between x and x_i (in electrostatic systems, $G(a, b) = 1/\left\|a - b\right\|_2$). From this definition, one can see that the potential at x is negative (positive) if x is in a neighborhood of relatively many negative (positive) charges. Thus, the potential indicates the sign and amount of charge in the local neighborhood.

Turning back to the ML classifier, one might propose a classification rule for some input x that assigns the label "+" if $\varphi(x) > 0$ or "−" otherwise. Abstracting from the electrostatic system, if $\alpha_i = 1$ and G is a function that decreases sufficiently steeply with distance, we obtain a nearest-neighbor classifier. (By "sufficiently steeply," we mean that if x_i is the closest point to x then $G(x_i, x) > N \, G(x_j, x) \, \forall j \neq i$.) The potential can also be viewed as the difference between a kernel density estimator for the "+" class and a kernel density estimator for the "$-$" class if $\alpha_i = \left|S^{y_i}\right|^{-1}$ ($S^{+1} \equiv S^+$ and $S^{-1} \equiv S^-$) and $\forall_a : \int G\left(a, x\right) dx = 1$.

1.2 Electrostatic model 2: Coupled point charges

Consider now an electrostatic model that extends the previous model in two respects. First, the point charges are replaced by *conductors*, e.g., metal spheres. Each conductor i has a *self-potential coefficient*, denoted P_{ii}, which is a measure of how much charge it can easily hold; for a metal sphere, P_{ii} is related to sphere's diameter. Second, the conductors in S^+ are *coupled*, as are the conductors in S^-. "Coupling" means that charge is free to flow between the conductors. (Technically, S^+ and S^- can each be viewed as a single conductor, but we will still use "conductor" in correspondence with $i \in \{1 \ldots N\}$.)

In this model, we initially place the same charge on each conductor, and allow charges within S^+ and S^- to flow freely (we assume no resistance in the coupling and no polarization of the conductors). After the charges redistribute, charge will tend to end up on the periphery of a homogeneous neighborhood of conductors, because like charges repel. Charge will also tend to end up along the S^+–S^- boundary because opposite charges attract. See Figure 1 for a depiction of the redistributed charges. The shading is proportional to the magnitude α_i.

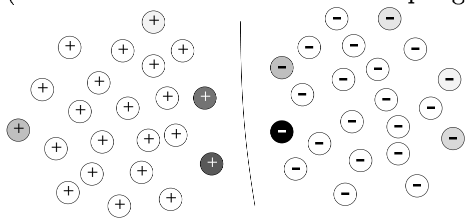

Figure 1: *Coupled conductor system at the energy minimum. Shading indicates the charge magnitude. The zero potential isoline is shown.*

An ML classifier can be built based on this model, once again using $\varphi(x) > 0$ as the decision rule for classification. In this model, however, the α_i are not uniform; the conductors with large α_i will have the greatest influence on the potential function. Consequently, one can think of α_i as the weight or importance of example i. As we will show shortly, the examples with $\alpha_i > 0$ are exactly support vectors of an SVM.

1.2.1 Formal Presentation

The potential on conductor i, $\varphi_i \equiv \varphi(x_i)$ can be described by the *coefficients of potential* P_{ij} [6]: $\varphi_i = \sum_{j=1}^{N} P_{ij}\, Q_j$, where $P_{ii} \geq P_{ij} \geq 0$ and $P_{ij} = P_{ji}$. P_{ij} specifies the potential induced on conductor i by charge Q_j on conductor j. To use a concrete physical example, if each conductor i is a metal sphere centered at x_i and has radius r_i, the system can be modeled by a point charge Q_i at x_i, $P_{ii} = G\left(x_i, \bar{x}_i\right)$, where $\bar{x}_i$ is an arbitrary point on the sphere surface, and $P_{ij} = G\left(x_i, x_j\right)$ [2, 6]. $G\left(a, b\right)$ must be isotropic, i.e., depend only on $\left\|a - b\right\|_2$. The free charge flow in S^+ and S^- corresponds to minimizing the Coulomb energy,

$$E = \frac{1}{2} \sum_{i=1}^{N} \varphi_i \, Q_i = \frac{1}{2} \, Q^T \, P \, Q = \frac{1}{2} \sum_{i,j=1}^{N} P_{ij} \, y_i \, y_j \, \alpha_i \, \alpha_j \; .$$

Initially, we set $\alpha_i = K/\left|S^{y_i}\right|$ to assign the same total charge magnitude K to S^+ and S^- and to make the charge uniform for each conductor in each set. Coulomb energy minimization redistributes the charges.

In order for this electrostatic model to serve as a classifier, we must enforce the constraint $\alpha_i \geq 0$ to ensure that an example does not change its class label. We do this by treating energy minimization as a constrained optimization problem with $0 \leq \alpha_i \leq C$, where C is an optional upper bound (which can be set to ∞ to eliminate the constraint). In the physical model, the constraint on α_i can be satisfied by disconnecting a conductor i from the charge flow in S^+ or S^- when α_i reaches the lower or upper bound, which will freeze its value.

After the energy minimum is reached, the potential will be the same for all $i \in S^+$ which are still connected; we denote this potential φ_{S^+}. Similarly, φ_{S^-} denotes the potential which is the same for all $i \in S^-$ which are still connected. To use the potential, $\varphi(x)$, to classify an input x, we must ensure that $\varphi_{S^+} = \varphi_{S^-}$ to eliminate any bias toward classification as "+" or "−". We can do so by introducing a constant potential b (something like ionized air in the physical system), i.e., $\varphi\left(x\right) = \sum_{i=1}^{N} Q_i \, G\left(x_i, x\right) + b$, where $b = -0.5\left(\varphi_{S^+} + \varphi_{S^-}\right)$.

We have described a system of coupled conductors with two additional constraints: (1) that the charge on a conductor is bounded, and (2) that positive and negative potentials are balanced. This physical system corresponds to a ν–support vector machine (ν–SVM) [5] if $C = 1/N$ and $\sum_{i \in S^+} \alpha_i = \sum_{i \in S^-} \alpha_i = 0.5 \; \nu$. The identity holds because the energy function is exactly the ν–SVM quadratic objective function, and in both the physical system and the SVM the function is minimized. We know from optimization theory that at the minimum, the Karush–Kuhn–Tucker conditions (KKTs) [1] must hold. The KKTs for ν–SVMs use the variables ρ, ξ_i, and μ_i which have a physical interpretation in our model. ρ is the potential difference between S^+ and S^-: $\rho = 0.5\left(\varphi_{S^+} - \varphi_{S^-}\right)$, or with b, we obtain $\rho = \pm\varphi_{S^\pm}$. Slack variable ξ_i gives the potential difference between φ_i and $\varphi_{S^{y_i}}$: $\xi_i = \rho - y_i \, \varphi_i \geq 0$. Removing conductors with $\alpha_i = 0$ from the system makes $\xi_i > 0$ only for $\alpha_i = C = 1/N$. Variable μ_i measures the charge difference to the upper bound $\mu_i = 1/N - \alpha_i \geq 0$ on i. The diagonal elements in the quadratic form have a physical interpretation as self–potential. As we discuss later, this interpretation will allow us to introduce novel kernels and novel SVM methods.

1.3 Electrostatic model 3: Coupled point charges with batteries

In electrostatic model 2, the same total charge is applied to S^+ and S^-

and the potentials $\varphi_{S^\pm}$ are balanced by b. However, we cannot control the magnitude of the potentials, $|\varphi_{S^\pm}|$. We can achieve this control by adding batteries to the system. We do this in two ways. In model 3.1, we connect S^+ to the positive pole of a battery with potential ϕ^+ and S^- to the negative pole with potential $\phi^- = -\phi^+$. The battery forces $\varphi_{S^+} = \phi^+$ and $\varphi_{S^-} = \phi^-$. The battery can then be removed and the potential remains. In model 3.2, we treat each conductor not as a (solid) sphere but as a spherical shell. We also connect each conductor shell i to its own battery, B_i, but not by direct contact. Rather, each shell i has a small sphere at its center which is connected to the positive pole of B_i if $y_i = -1$ and the negative pole if $y_i = +1$ (Figure 2). Consequently, the induced constant potential, ϕ_i, has polarity opposite that of the conductor $(-y_i)$. To add charges to S^+ and S^- we ground both. Charges flow into the system until the potentials equalize. Therefore, after removing the batteries and fixing the charges we have $\varphi_i = -\phi_i$ (unless a conductor is disconnected).

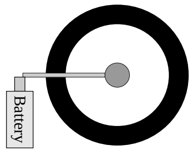

Figure 2: *Conductor with battery.*

1.3.1 Formal Presentation

$\phi_i = -\ \beta_i\ y_i$ $(\beta_i \geq 0)$ is the potential induced by the battery B_i on conductor i, the total potential on conductor i is $\Phi_i = \varphi_i + \phi_i$, the energy contribution of the battery B_i is $1/2\ \phi_i\ Q_i$ [2], and the total Coulomb energy is:

$$\frac{1}{2}\sum_{i=1}^{N}(\Phi_i+\phi_i)\;Q_i=\frac{1}{2}\,Q^T\,P\,Q\;+\;\phi^T\,Q\;=\frac{1}{2}\sum_{i,j=1}^{N}P_{ij}\,y_i\,y_j\,\alpha_i\,\alpha_j\;-\;\sum_{i=1}^{N}\beta_i\,\alpha_i\;.$$

This physical system corresponds to a C–support vector machine (C–SVM) [1, 8] if $\forall i:\ \beta_i = 1$ (that is, model 3.1 with $|\phi^\pm| = 1$). The Coulomb energy is the C–SVM objective function. Our model yields $\varphi_{S^+} = -\varphi_{S^-}$; consequently, we do not require b from model 2 and the C–SVM constraint $\sum_i y_i\alpha_i = 0$ is not necessary. The KKT-condition variables receive a physical interpretation analogous to that in the ν–SVM.

2 Comparison of existing and novel models

2.1 Novel Kernels

$E\ =\ \int G\,(x,y)\ \ h\,(x)\ \ h\,(y)\ \ dxdy\ \ \geq\ \ 0$ must hold in a continuous physical system for the energy E. Here h^+ (h^-) is the density of positive (negative) charges and $h = h^+ - h^-$. This is exactly Mercer's condition in the context of SVM which ensures positive definite kernels [1]. To maintain properties of the physical model (e.g., $b = 0$ in model 3.2), we fulfill Mercer's condition by restricting G to isotropic kernels, i.e., $G\,(x_i, x_j)\ \equiv\ g(\|x_i - x_j\|_2^2)$, where g is completely monotonic, i.e., $(-1)^k\ g^{(k)}\ (x) \geq 0$, $\forall x \geq 0$ [7].

The electrostatic perspective makes apparent that SVM algorithms can break down in high dimensions. The reason is that fast decreasing kernels induce small potentials and, therefore, almost every conductor retains charge. We want to use kernels which do not decrease exponentially. The self–potential allows the use of kernels that would otherwise be invalid, such as a generalization of the electric field to d dimensions: $g\left(z\right) = z^{1-0.5d}$, where we define $G\left(x_i, x_i\right) := P_{ii} = g\left(r_i^2\right)$. Smoothing this kernel by ϵ and using an exponent n leads to the Plummer potential which is used in computational physics to simulate electrostatic fields $g\left(z\right) = \left(z + \epsilon^2\right)^{-0.5n}$ with $r_i = \min_j \left\|x_i, x_j\right\|_2$. For $c \geq c_0 = \max\{0.5\ z \mid z = \left\|x_i - x_j\right\|_2^2 \ \vee \ z = r_i^2\}$ (we used $c = c_0$) is $g\left(z\right) = \left(c - 0.5z\right)^n$ a polynomial and for $n = 1$ the conventional linear kernel.

2.2 Novel SVM models

Our electrostatic framework can help to derive many distinct SVM approaches, several representative examples we now illustrate.

2.2.1 κ–Support Vector Machine (κ–SVM)

We can exploit the physical interpretation of P_{ii} as conductor i's self–potential, (i.e., how easy it is to put charges on i). The P_{ii}'s determine the entropy of the charge distribution at the energy minimum. We can rescale the self potential—$P_{ii}^{new} = \kappa\, P_{ii}^{old}$—and use κ to control the complexity of the SVM in electrostatic models 3.1 and 3.2 with $C = \infty$.

2.2.2 p–Support Vector Machine (p–SVM)

Without constraints, $PQ + \phi = 0$ at the energy minimum of model 3.1 and 3.2, which is $\forall i : \ \varphi_i + \phi_i = 0$. In physical terms this means that *potentials equalize.* However, the solution $Q = -P^{-1}\phi$ suffers from violating the constraint that $\alpha_i \geq 0$. We can instead minimize the potential difference, $\frac{1}{2}\left\|PQ + \phi\right\|_2^2 = \frac{1}{2}Q^T P^T P Q + Q^T P^T \phi + \frac{1}{2}\phi^T\phi$, where the last term is constant. Without constraints, the minimum is $Q = \left(P^T P\right)^{-1} P^T \phi$, where $\left(P^T P\right)^{-1} P^T$ is P's pseudo inverse. Using physical model 3.1, and defining $\mu_i := \sum_{j=1}^{N} y_i y_j P_{ij}$, we obtain:
$\min_\alpha \ \frac{1}{2}\alpha^T K \alpha - \mu^T \alpha \ \ \text{s.t.} \ \ y^T\alpha = 0 \wedge 0 \leq \alpha_i \leq C$, where $K_{ij} := y_i y_j \left[P^T P\right]_{ij}$. K is by construction positive definite so that **this formulation does not demand positive definite kernels.** If we set $\beta_i = 1/\mu_i$ then we obtain the generalized SVM in [3]; however, for other values of β_i (e.g., $\beta_i = 1$) we obtain an SVM that automatically removes outliers, e.g., the p–SVM. Outliers gets a negative or small μ_i, which results in a small α_i.

2.3 Experiments

For the representative models we've introduced, we perform simulations and make comparisons to the standard SVM models. The datasets are from the UCI Benchmark Repository and preprocessed in [4], where the "banana" data set stems from (`http://www.first.gmd.de/~raetsch/data`). We did 100-fold validation on each data set, restricting the training set to 200 examples, and using the remainder of examples for testing. We compared C–SVM, ν–SVM, κ–SVM, and p–SVM. Additionally we combined the later to κ–p–SVM allowing κ values which lead to not positive definite kernels. We used radial basis function (RBF), polynomial (POL), and Plummer (PLU) kernels. Hyperparameters are determined by 5–fold cross validation on the first 5 training sets. The search for hyperparameter was not as intensive as in [4].

	C	ν	κ	p	κ-p	C	ν	κ	p	κ-p
	thyroid					heart				
RBF	6.4	9.4	7.7	**5.4**	8.6	21.4	19.1	17.9	22.4	*17.8*
POL	22.8	12.6	7.0	13.3	6.9	20.4	20.4	19.3	23.0	19.3
PLU	*6.1*	6.2	*6.1*	*5.7*	*6.1*	**16.3**	**16.3**	**16.3**	*17.4*	**16.3**
	breast–cancer					banana				
RBF	33.6	31.6	33.8	32.4	33.7	*13.2*	36.7	*13.2*	*11.6*	13.4
POL	36.0	**25.7**	29.6	*27.1*	*29.1*	35.3	35.0	**11.5**	22.4	**11.5**
PLU	33.4	33.1	33.4	30.6	33.4	15.7	15.7	15.7	21.9	15.7
	german									
RBF	28.7	29.3	29.0	*27.8*	28.8					
POL	33.7	29.6	**26.2**	31.8	**26.2**					
PLU	28.8	28.5	33.3	*27.1*	33.3					

Table 1: *Mean % misclassification over 100 replications. The columns correspond to SVMs and the rows to kernel functions.*

The Plummer potential is more robust against hyperparameter and SVM choices. The proposed novel methods performed well compared to known approaches.

2.4 Other SVM approaches

This work leads to many models that could be explored. For example, the variables β_i in model 3.2 were not further investigated. With fixed charge, β_i determines how conductor i retains its charge. Here, however, we will present SVM speed ups.

2.4.1 Support Vector Machine By Linear Programming

We minimize $||P\ Q\ +\ \phi||_1$ by minimizing $\sum_{i=1}^{N} s_i$ with constraints $\beta_i - s_i \leq y_i\ [P\ Q]_i \leq \beta_i + s_i$, $\sum_i y_i\ \alpha_i = 0$, and $\alpha_i \geq 0$. Maximizing the β_i as well results in the linear SVM formulation, e.g., [3].

2.4.2 Support Vector Machine By Solving One Equation

We will adjust the P_{ii} so that $Q = -P^{-1}\phi$ does not violate $\alpha_i \geq 0$. We divide $P = \tilde{P} + D$ into diagonal matrix D $(D_{ii} = \kappa_i)$ and zero diagonal matrix $\tilde{P}$.

Fast, iterative algorithm.

$\kappa_i \geq \sum_{j,j\neq i} P_{ij}$ ensures $\alpha_i \geq 0$. This means that P is diagonal dominant and the fast Jacobi iteration is possible.

Standard equation solving algorithms.

We set $\forall_i\ :\ \kappa_i = \kappa_0$ and perform a k–step bisection to find a minimal κ_0 which does not violate $\alpha_i \geq 0$.

2.4.3 Support Vector Machine By A Quick and Dirty Approximation

We solve $y_i \sum_{j=1}^{N} y_j P_{ij} \alpha_j\ =\ \beta_i$ with the assumption that conductors are surrounded by conductors with the same charge magnitude, i.e. $\alpha_j\ = \alpha_i$. We get $\alpha_i = \beta_i / \mu_i$, where we keep $\mu_i \geq \epsilon$.

2.5 Vector quantization and clustering

SVMs focus on the boundaries whereas vector quantization and clustering algorithms focus on high density regions in order to obtain prototype vectors or cluster centers. This corresponds to energy maximization in our physical systems with $\alpha_i \geq \epsilon$. We get a dual between SVM and vector quantization/clustering. For example, constraints can determine the number of clusters or prototypes.

3 Conclusion

The electrostatic framework and its analogy to SVMs has led to several important ideas: (1) It suggests SVM methods that are valid for kernels that are not positive definite. (2) It allowed us to derive fast SVM methods based on linear programming and linear equations. (3) It suggested novel approaches and kernels that perform at least as well as standard methods.

We argued that the electrostatic framework not only characterizes a family of support-vector machines, but it also characterizes other techniques such as nearest neighbor classification, classification by density estimation, vector quantization, and clustering. Perhaps the most important contribution of the electrostatic framework is that, by interrelating and encompassing a variety of methods, it lays out a broad space of possible algorithms. At present, the space is sparsely populated and has barely been explored. But by making the dimensions of this space explicit, the electrostatic framework allows one to easily explore the space and discover novel algorithms. In the history of machine learning, such general frameworks have led to important advances in the field.

References

[1] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):1–47, 1998.

[2] L. N. Kantorovich, A. I. Livshits, and M. Stoneham. Electrostatic energy calculation for the interpretation of scanning probe microscopy experiments. *Journal of Physics: Condensed Matter*, 12:795–814, 2000.

[3] O. Mangasarian. Generalized support vector machines. Technical Report 98-14, Computer Sciences Dep., Univ. of Wisconsin, Madison, Wisconsin, 1998.

[4] G. Rätsch, T. Onoda, and K.-R. Müller. Soft margins for AdaBoost. Technical Report NC-TR-1998-021, Dep. of Computer Science, Univ. of London, 1998.

[5] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000.

[6] M. Schwartz. *Principles of Electrodynamics*. Dover Publications, NY, 1987.

[7] A. J. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularitzation operators and support vector kernels. *Neu. Net.*, 11:211–231, 1998.

[8] V. Vapnik. *The nature of statistical learning theory*. Springer, NY, 1995.