Mach Learn (2008) 72: 173–188
DOI 10.1007/s10994-008-5071-9

Large margin vs. large volume in transductive learning

Ran El-Yaniv · Dmitry Pechyony · Vladimir Vapnik

Received: 23 June 2008 / Revised: 23 June 2008 / Accepted: 23 June 2008 / Published online: 8 July 2008

Abstract We consider a large volume principle for transductive learning that prioritizes the transductive equivalence classes according to the volume they occupy in hypothesis space. We approximate volume maximization using a geometric interpretation of the hypothesis space. The resulting algorithm is defined via a non-convex optimization problem that can still be solved exactly and efficiently. We provide a bound on the test error of the algorithm and compare it to transductive SVM (TSVM) using 31 datasets.

Keywords Transductive learning · Large margin · Large volume · TSVM · Learning principles

1 Introduction

Alternative learning models that utilize unlabeled data have received considerable attention in the past few years. Two prominent models are semi-supervised and transductive learning. The main attraction of these models is empirical evidence indicating that they can often allow for more efficient and significantly faster learning in terms of sample complexity.

In this paper we focus on distribution-free *transductive learning*. In this setting the learning algorithm is given a fixed individual set of unlabeled data points whose labels are hidden. Then, a training sample is selected randomly from the set and the labels of the training points are revealed. The goal is to predict the labels of the remaining unlabeled points as accurately as possible.

Editors: Walter Daelemans, Bart Goethals, Katharina Morik.

R. El-Yaniv · D. Pechyony (✉)
Computer Science Department, Technion-Israel Institute of Technology, Haifa 32000, Israel
e-mail: pechyony@cs.technion.ac.il

R. El-Yaniv
e-mail: rani@cs.technion.ac.il

V. Vapnik
NEC Laboratories America, Princeton, NJ 08540, USA
e-mail: vapnik@att.net

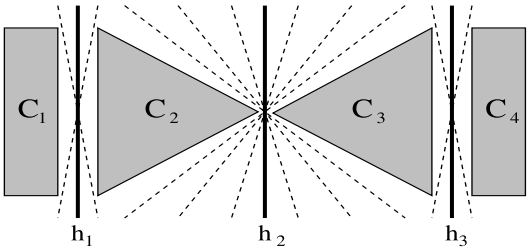

Fig. 1 Large-margin vs. large-volume prior

In both transduction and semi-supervised induction we would like to benefit from additional unlabeled data to improve learning rates and accuracy. In semi-supervised induction the additional unlabeled data may provide useful information on the marginal unknown distribution $p(x)$ of unlabeled samples x. In distribution-free transduction, $p(x)$ is already fully known. In this regard, the transductive learning problem is easier. Moreover, in transduction the given test examples provide an additional advantage that can be appreciated from a learning theoretic perspective, which views learning as a *selection* process of a hypothesis from an hypothesis class. In this perspective, the conceptual advantage of transduction over (semi-supervised) induction is the possibility to perform data dependent selection and ranking of candidate hypotheses *after* observing the unlabeled data. This, in turn, allows us to fix our attention to smaller hypothesis spaces that are more relevant to the data at hand.

In transductive binary classification, any hypothesis space (say, hyperplanes) is reduced to a finite collection of equivalence classes, given the unlabeled data. All hypotheses in the same class are identical in their binary classification of the data. For example, consider Fig. 1, in which $C_1, \ldots, C_4$ represent four "clouds" of unlabeled data. In this case all the hyperplanes passing in between C_3 and C_4 (and, in general, between C_i and C_{i+1}) are in the same equivalence class (as well as infinitely many other hyperplanes that are not shown). The extra advantage in transduction is the possibility to prioritize these equivalence classes in accordance with our prior beliefs about the goodness of hypotheses given the current data. A classic principle for prioritizing equivalence classes is the large margin principle introduced in (Vapnik 1982). According to this principle, the priority (or prior) of an equivalence class is proportional to the largest margin obtained by any of its members. In our example in Fig. 1 we should prefer the equivalence classes of $\mathbf{h}_1$ and $\mathbf{h}_3$ over the class of $\mathbf{h}_2$ because they achieve a larger margin. This large margin consideration motivated the transductive support vector machine (TSVM) (Vapnik 1998) approach for transduction.

In this paper we consider a different, *large volume* principle, whereby the prior of an equivalence class is proportional to its "volume" in the hypothesis space. For example, in the case of hyperplanes, in Fig. 1 we should prefer the equivalence class of $\mathbf{h}_2$ because it has a much larger volume in the hyperplane space. This is depicted in Fig. 1 by the number of dashed lines that pass between the clouds.

The large volume transductive principle was briefly treated in (Vapnik 1982) for the case of hyperplanes. Here we further study this principle and instead of hyperplane spaces we consider general soft classification vectors whose set of equivalence classes with respect to all data points (ignoring labels) contains all possible dichotomies. Symmetry is broken by generating equivalence classes of non-uniform volume, defined via a non axis aligned data-dependent ellipsoid. Since exact or quantifiable volume approximation is computationally hard, we resort to a cruder approach whereby we measure the angles between hypotheses to the principal axes of the ellipsoid. This approach makes sense because long principal axes lie in regions of large volume. This construction leads to a general family of transductive algorithms and here we focus on one instantiation. Although the resulting algorithm is defined

in terms of a non-convex optimization problem, we develop an efficient global optimum solution using a known technique. We also derive a transductive data-dependent error bound for this algorithm.

Our empirical evaluation of the new algorithm over a large number of datasets shows its overwhelming advantage over TSVM (and SVM) in text categorization and image classification problems. However, on a different set of UCI datasets, TSVM and SVM are significantly superior to the new algorithm. In our analysis of this finding, we identify some factors that influence the success and failure of our algorithm. In particular we show that our algorithm has significant advantage over TSVM when TSVM outperforms SVM.

2 The transductive setting

We consider the following distribution-free transductive model (Vapnik 1998, p. 341, Setting 1). A fixed set $X_{m+u} = \{x_1, \ldots, x_{m+u}\}$ of $m+u$ points from some space $\mathcal{X}$ is given. The binary labels $y_i \in \{\pm 1\}$ of these points are also fixed but unknown to us. We uniformly at random pick a subset $X_m \subseteq X_{m+u}$ of size m (among all subsets of size m), and the labels of these points are provided. Rearranging indices, we denote by $S_m = \{(x_i, y_i)\}_{i=1}^{m}$ the given labeled points, and by $X_u = \{x_j\}_{j=m+1}^{m+u}$, the remaining u unlabeled points. Using S_m and X_u, our goal is to guess the labels of points in X_u as accurately as possible.

Fixing m and u, we consider soft "hypotheses" that are *vectors*

$$\mathbf{h} = (h_1, \ldots h_{m+u}) \in \mathbb{R}^{m+u},$$

where h_i is the soft, or confidence-rated label of example x_i given by "hypothesis" $\mathbf{h}$. The vector $\mathbf{h}$ can be also interpreted as a *functional response vector* w.r.t. some underlying function f such that for any $1 \le i \le m+u$, $h_i = f(x_i)$. Based on knowledge of the full-sample X_{m+u}, the learning algorithms we consider select an hypothesis space $\mathcal{H} = \mathcal{H}(X_{m+u})$ of such soft classification vectors. Then, given the labels of training points the algorithm selects one hypothesis $\mathbf{h} \in \mathcal{H}$. For actual (binary) classification of x_i the algorithm outputs $\mathrm{sgn}(h_i)$.

Two quantities of interest, for an hypothesis $\mathbf{h}$, are its transductive risk, or *test error*, $R_u^\ell(\mathbf{h}) \triangleq \frac{1}{u}\sum_{i=m+1}^{m+u} \ell(h_i, y_i)$, defined w.r.t. some loss function ℓ, and the training or *empirical error* (w.r.t. ℓ), $R_m^\ell(\mathbf{h}) \triangleq \frac{1}{m}\sum_{i=1}^{m} \ell(h_i, y_i)$. In this paper ℓ will be instantiated to the zero-one loss, the hinge loss, and linear loss functions. Whenever we omit ℓ from R_u^ℓ and R_m^ℓ, we assume that the zero-one loss function is used.

3 Transductive maximum power inference

Let $\mathcal{H}$ be any (soft) hypothesis space. A crucial observation, made by Vapnik (1982) for a classification setting, is that after the introduction of the unlabeled data X_{m+u}, the set $\mathcal{H}$ is partitioned into a finite number of *equivalence classes* $H_1, \ldots, H_N$, such that all hypotheses in $H_k, k = 1, \ldots, N$, generate the same dichotomy of X_{m+u}. Suppose that there exists some underlying distribution $P(\mathbf{h})$ over $\mathcal{H}$ such that one hypothesis is selected randomly according to P and the selected hypothesis determines the labels of points in X_{m+u}. Vapnik (1998, p. 708) defined the *power* of an equivalence class H_k as the probability mass (in terms of P) of all the soft hypotheses in it,

$$\mathrm{Power}(H_k) \triangleq \int_{H_k} dP(\mathbf{h}), \quad k = 1, \ldots, N. \tag{1}$$

The power function provides a preference relation over all the dichotomies of X_{m+u} that can be generated by $\mathcal{H}$. So, for example, if we utilize an empirical error minimization framework, then, among all equivalence classes that classify the training set correctly, we should prefer one that has the largest power. We term this principle (that was already proposed in Vapnik 1998, pp. 707–708) '*transductive maximum power inference*'. The principle can also be extended to structural risk minimization.

In practice, of course, we do not know the underlying hypothesis distribution (and moreover, such a distribution may not exist) so in order to implement maximum power inference we must make a guess about some *prior* distribution P over $\mathcal{H}$, or directly infer $\text{Power}(H_k)$ for $k = 1, \ldots, N$. Obviously, a good prior distribution P should reflect auxiliary knowledge on the effectiveness of soft hypotheses.

If the power function is only dependent on the unlabeled data (and not on the train/test partition and the labels), the following error bound, which is an immediate consequence of Theorem 22 in (Derbeko et al. 2004), provides a compelling motivation for maximum power inference: for any $0 < \delta < 1$, with probability of at least $1 - \delta$ over choices of S_m, for all $k = 1, \ldots, N$,

$$R_u(H_k) \leq R_m(H_k) + \sqrt{\left(\ln \frac{1}{\text{Power}(H_k)} + \ln \frac{1}{\delta}\right) \cdot \left(\frac{1}{m} + \frac{1}{u}\right)}, \tag{2}$$

where $R_m(H_k)$ (respectively $R_u(H_k)$) is the training (respectively test) error obtained by any instance of H_k. The error bound (2) implies that if an equivalence class H_k with a large power is empirically successful (over the training set), its test error over X_u is guaranteed to be small, with high probability.

4 On priors and powers

The bound (2), which essentially provides a sufficient condition for transductive learning, tells us that the power of the equivalence classes is a crucial quantity that can directly affect the learning speed and accuracy. Power assignment can be based on 'low-level' considerations, via prior assignment for hypotheses, as in (1). However, this assignment can also be done directly on complete equivalence classes, without defining a prior distribution P over soft hypotheses. In the latter case, the power is simply a prior over equivalence classes.

Various approaches of defining prior directly over equivalence classes have been considered in the past. The most well known approach is the *maximum margin* principle given by (Vapnik 1982). The margin of a hyperplane is its minimal distance to any example in the full sample. By the maximum margin principle, the prior of the equivalence class H_k is proportional to the maximal margin obtained by any hyperplane $h \in H_k$. The maximal-margin principle motivated the well known transductive SVM (TSVM) approach. Other prior assignment approaches using compression, clustering and graph cuts are discussed in (Derbeko et al. 2004) and (Hanneke 2006).

Effective power assignments must rely on some specialized knowledge that requires insight into the learning problem at hand. For some problems, priors on soft hypotheses (or power of equivalence classes) can be difficult to identify or quantify. Vapnik proposed an alternative prior encoding scheme through the *universum* (Vapnik 1998, p. 707). The universum is a set of unlabeled examples belonging to the same domain $\mathcal{X}$, but are known not to belong to any one of the two classes. The power of an equivalence class should be

taken as the number of dichotomies it obtains over the universum examples.[1] Since counting the number of dichotomies is computationally hard, it was proposed to approximate it with the number of contradictions. A universum example x contradicts an equivalence class H_k if there exists a pair of soft hypotheses $\mathbf{h}, \mathbf{h}' \in H_k$, such that $h(x) \neq h'(x)$. We term this approximation as the (universum) *maximum contradiction* principle.

Although the universum approach as presented above is transductive, it can also be motivated for induction. In fact, the first empirical study and validation of the universum idea is within an inductive setting (Weston et al. 2006), where an hypothesis class of hyperplanes is considered and it is suggested to approximate the number of contradictions of an equivalence class H_k by the minimum, over all $\mathbf{h} \in H_k$, of the sum of ℓ_1-distances of $\mathbf{h}$ from all universum examples. The intuition behind this approximation is that a very close proximity of h to a universum example x implies that a slight perturbation in the direction of $\mathbf{h}$ will generate a new $\mathbf{h}'$ that classifies x differently. The success of this approximated maximum contradictions principle depends on the choice of universum examples and it was shown in (Weston et al. 2006) that a combination of both the maximum margin and the maximum contradictions principles can outperform the maximum margin principle alone, if the universum is effectively selected.

In some domains universum examples arise naturally. For example, in a binary recognition problem where we want to separate the digits '1' and '2', examples of other digits can form an effective universum (Weston et al. 2006). But in general, universum examples may be hard to generate, especially in problems where we cannot easily perceive the membership of the universum examples to the domain.

5 A large volume principle

Consider a transductive classification setting and assume for now that $\mathcal{H}$ (which may depend on X_{m+u}) is finite. We consider the assignment of a prior measure P over $\mathcal{H}$. In the absence of any other knowledge, by the principle of insufficient reason, the prior of *any* two soft hypotheses (not necessarily from the same equivalence class) should be the same. This, of course, does not imply that the powers of two equivalence classes are identical.[2] According to (1), if P is uniform and $\mathcal{H}$ is finite then the power of any equivalence class is proportional to its size. A straightforward extension of this argument to a continuously infinite (soft) $\mathcal{H}$ results in a power function that assigns to each equivalence class the geometric volume of soft hypotheses contained in it. We term this application of the maximal power inference principle with a uniform prior (over the soft hypotheses) the *large volume* principle.

There are a few previous works that explicitly or implicitly utilized a large volume principle for an hypothesis space of (kernelized) hyperplanes. Vapnik (1982, Sect. 10.5) proposed to approximate the volume of an equivalence class (of hyperplanes) by the distance between convex hulls of positive and negative examples. As shown by (Bennett and Bredensteiner 2000), this distance is precisely the margin.

Tong and Koller (2001) exploited a duality between hyperplanes and instance points, where hyperplanes are viewed as points on a sphere and examples are viewed as hyperplanes passing through the sphere. They approximated the volume of an equivalence class

[1] In philosophical terms, the universum is used to measure *falsifiability* (or *specificity*)—the more powerful equivalence classes are those that are more *falsifiable* by the universum points (Vapnik 1998).

[2] Note that whenever the number of equivalence classes is $\Omega(2^{m+u})$, if the power is uniform over classes, we cannot bound the transductive test error.

(corresponding to the version space) by the radius of the maximally inscribed ball within a conic section. This radius is precisely the margin and the approximation can be arbitrarily poor whenever the equivalence class is an elongated section.

Again for hyperplanes, Graepel et al. (1999) approximated the volume of hypothesis equivalence classes using a kernel billiard algorithm. Their algorithm operates in a transductive setting, but considers equivalence classes defined by training points and a single test point. In contrast, we consider here equivalence classes defined by all training and test points.

Finally, we observe that one can approximate the volume using uniformly drawn universum examples. In this case one can show that, asymptotically, the equivalence classes with larger volume will have a larger number of contradictions.[3]

The main difference between our contribution and the previous work described above is that instead of hyperplane spaces we consider a much richer space of general soft classification vectors. This space, unlike hyperplanes, generates all possible 2^{m+u} dichotomies.

6 Transductive learning using the large volume principle

We describe a transductive learning scheme that approximates the large volume principle. This scheme motivates a family of new transductive algorithms. In this section we develop and analyze one instantiation of this scheme.

6.1 Volume approximation

Our approach for approximating the volume of the equivalence classes relies on hypothesis spaces that can be represented as ellipsoids in $\mathbb{R}^{m+u}$. Each soft hypothesis in the hypothesis space is a point in the ellipsoid. We approximate the volume of an equivalence class H_k by the angles between an (arbitrary) hypothesis in H_k and the principal axes of the ellipsoid.

Let the full sample X_{m+u} be given and fixed. Let $\mathbf{h} \in \mathbb{R}^{m+u}$ be a soft transductive hypothesis, and Q, a positive definite $(m+u) \times (m+u)$ matrix that may depend on X_{m+u}. The matrix Q is considered as a hyperparameter and in Sect. 9 we give an example for its instantiation. Consider a hypothesis space $\mathcal{H}_Q = \{\mathbf{h} \mid \mathbf{h}^T Q \mathbf{h} \leq 1\}$. Geometrically, $\mathcal{H}_Q$ is an ellipsoid in $\mathbb{R}^{m+u}$, centered at zero. We denote it by $\mathcal{E}(\mathcal{H}_Q)$. Since Q is positive definite, the set $\{\text{sign}(\mathbf{h}) : \mathbf{h} \in \mathcal{H}_Q\}$ contains all 2^{m+u} possible dichotomies of X_{m+u}.

The Cartesian coordinate system divides the space $\mathbb{R}^{m+u}$ into 2^{m+u} quadrants. Each quadrant corresponds to one equivalence class in terms of hard classification. For any $1 \leq k \leq 2^{m+u}$, the volume of the equivalence class H_k is the volume of the intersection of $\mathcal{E}(\mathcal{H}_Q)$ with the kth quadrant. For example, Fig. 2(a) shows four equivalence classes, $H_k, k = 1, \ldots, 4$, for the case $m + u = 2$. Ultimately, we would like to compute the exact volume of these quadrant intersections. However, currently known algorithms for approximate volume computation of general convex bodies seem to be too slow for practical purposes (Lovasz and Vempala 2006).

We resort to the following heuristic approximation. Let $\{(\lambda_i, \mathbf{v}_i)\}_{i=1}^{m+u}$ be the eigenvalues of Q along with their eigenvectors, such that for all $2 \leq i \leq m+u$, $\lambda_{i-1} \leq \lambda_i$. We assume

[3] Proof outline: consider a dual space of hyperplanes with the offset $b = 0$ (w.l.o.g.). This space is a sphere and full sample points are hyperplanes passing through the origin and cutting the sphere. Each equivalence class is a conical section of this sphere. In the dual space, uniformly drawn universum examples are equivalent to uniformly drawn hyperplanes. Thus, a universum example generates a contradiction in a conical section iff its hyperplane cuts the section. If the conical section is large then it will be cut by many hyperplanes.

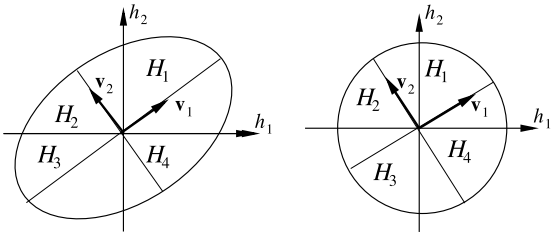

Fig. 2 Visualization of hypothesis space: (**a**) equivalence classes have different volumes; (**b**) equivalence classes have the same volume

w.l.o.g. that for any $1 \le i \le m+u$, $\|\mathbf{v}_i\|_2 = 1$. The direction and length of the ith principal axis of $\mathcal{E}(\mathcal{H}_Q)$ are, respectively, $\mathbf{v}_i$ and $\sqrt{1/\lambda_i}$. As shown in Fig. 2(a), the volume of H_k is proportional to the length of the principal axes of the ellipsoid, which falls in its quadrant. In the extreme case of a perfect sphere (Fig. 2(b)), all equivalence classes are of the same volume and cannot be differentiated. Therefore, we should prefer skewed ellipsoids that ultimately reflect useful priors on hypothesis effectiveness. In Sect. 9.1 we give example of such skewed ellipsoids that yield preference for "smoother" hypotheses.

The number of principal axes is always $m+u$ whereas the number of quadrants (and equivalent classes) is 2^{m+u}. Hence, the vast majority of quadrants do not contain any principal axis and, unlike the 2-dimensional case, we cannot estimate the volume of an equivalence class using a corresponding principal axis. We propose to estimate the volume using a weighted sum of axes' lengths such that the weights are determined by the polar proximity of an hypothesis to the principal axes.

Fix i and j such that $1 \le i < j \le m+u$ and $\lambda_i < \lambda_j$. Then, the length of the ith principal axis is larger than the length of the jth one. Hence, the local neighborhood of $\mathbf{v}_i$ has a larger volume than that of $\mathbf{v}_j$. A small angle between an hypothesis $\mathbf{h}$ and some long principal axis is taken as an evidence that its equivalence class has large volume. Conversely, a small angle to a short principal axis is taken as an evidence of a small volume. Note that these two opposing conditions cannot be satisfied simultaneously since the principal axes are orthogonal. In Sect. 9 we briefly discuss the meaning of the eigenvectors for a particular Q of interest.

Let $0 \le a_1 \le a_2 \le \cdots \le a_{m+u}$ be an increasing sequence of weights. For any soft hypothesis $\mathbf{h}$ let

$$V(\mathbf{h}) = \sum_{i=1}^{m+u} a_i \frac{(\mathbf{h}^T \mathbf{v}_i)^2}{\|\mathbf{h}\|_2^2}. \tag{3}$$

The expression $(\mathbf{h}^T \mathbf{v}_i)^2 / \|\mathbf{h}\|_2^2$ is the square of the cosine of the angle between $\mathbf{h}$ and the unit-length vector $\mathbf{v}_i$. The monotone increasing sequence of a_i's corresponds to a monotone decreasing sequence of the lengths of $\mathbf{v}_i$'s. Thus, the weighted sum (3) gives larger weight to the angular closeness to short principal axes than to the long ones. Consequently, we expect (3) to be large when $\mathbf{h}$ lies in the equivalence class of *low* volume and be small when $\mathbf{h}$ lies in the equivalence class of *high* volume.

6.2 Approximate volume regularization (AVR) algorithm

We propose the following natural instantiation of the $\{a_i\}_{i=1}^{m+u}$ such that they are inversely proportional to the lengths of their corresponding principal axes. Let $d_i = \sqrt{1/\lambda_i}$ be the length of the ith largest principal axis of the ellipsoid and for any $\mathbf{h} \in \mathbb{R}^{m+u}$, set $a_i = 1/d_i^2 =$

λ_i. Then,

$$V(\mathbf{h}) = \sum_{i=1}^{m+u} \lambda_i \frac{(\mathbf{h}^T \mathbf{v}_i)^2}{\|\mathbf{h}\|_2^2} = \frac{\mathbf{h}^T Q \mathbf{h}}{\|\mathbf{h}\|_2^2}.$$

This volume approximation motivates the following family of transductive algorithms, which implements the large volume principle:

$$\min_{\mathbf{h} \in \mathcal{H}_Q} R_m(\mathbf{h}) + \gamma \cdot \frac{\mathbf{h}^T Q \mathbf{h}}{\|\mathbf{h}\|_2^2}, \tag{4}$$

where $\gamma > 0$ is a regularization parameter.[4]

Instead of the 0/1 loss empirical error in (4), due to computational considerations (see Remark 3 in Sect. 9), we instantiate the loss function to the linear loss, $\ell(h_i, y_i) \triangleq -h_i y_i$. Fixing $t > 0$ and constraining $\mathbf{h}$ to be of length t we eliminate the denominator in (4). Also, we replace the constraint $\mathbf{h} \in \mathcal{H}_Q$ with $\mathbf{h} \in \mathbb{R}^{m+u}$ (see below). The resulting optimization problem is

$$\min_{\mathbf{h} \in \mathbb{R}^{m+u}} \quad -\frac{1}{m} \sum_{i=1}^{m} h_i y_i + \gamma \cdot \mathbf{h}^T Q \mathbf{h} \tag{5}$$

$$\text{s.t.} \qquad \|\mathbf{h}\|_2^2 = t^2. \tag{6}$$

We refer to the optimization problem (5)–(6) as the *Approximate Volume Regularization (AVR) algorithm*. Due to constraint (6) the loss $-h_i y_i$ of each training example is lower bounded by $-t$. Notice that while the optimization in (5) is done in $\mathbb{R}^{m+u}$, the regularization is done relative to $\mathcal{H}_Q$. The reason is that under the constraint (6) the complexity term $\mathbf{h}^T Q \mathbf{h}$ is a weighted sum of the squares of cosines between $\mathbf{h}$ and the principal axes of $\mathcal{E}(\mathcal{H}_Q)$. Thus, the optimization problem (5)–(6) is directly implied by (4) under the above instantiations of the free parameters.

While the optimization problem (5)–(6) is not convex, it can be solved efficiently and exactly (to obtain a global optimum) using the method of (Gander et al. 1989). This solution is developed in Sect. 7.

7 Global optimum AVR optimization

Following (Gander et al. 1989), we solve (5)–(6) using Lagrange multipliers. Set

$$\Phi(\mathbf{h}, \rho) = -\frac{1}{m} \sum_{i=1}^{m} h_i y_i + \gamma \cdot \mathbf{h}^T Q \mathbf{h} - \rho(\|\mathbf{h}\|_2^2 - t^2),$$

where ρ is a Lagrange multiplier. Then, $\mathbf{h}^* = \min_{\mathbf{h} \in \mathbb{R}^{m+u}, \rho \in \mathbb{R}} \Phi(\mathbf{h}, \rho)$ is a solution of (5)–(6). The minimum of $\Phi(\mathbf{h}, \rho)$ is achieved when its partial derivatives are zero. Let $\mathbf{y} \in \mathbb{R}^{m+u}$ be a vector of labels, with the first m entries being the training labels and the last u entries being

[4]Note that one deficiency of the above approximation is that for two hypotheses $\mathbf{h}$ and $\mathbf{h}'$ from the same equivalence class, $V(\mathbf{h})$ is not in general identical to $V(\mathbf{h}')$.

zeros. Differentiating $\Phi(\mathbf{h}, \rho)$ w.r.t. $\mathbf{h}$ and ρ, and equating both these derivatives to zero we get,

$$-\mathbf{y}/m + 2\gamma Q\mathbf{h} - 2\rho\mathbf{h} = 0; \tag{7}$$

$$\|\mathbf{h}\|_2^2 - t^2 = 0. \tag{8}$$

It follows from (7) that[5]

$$\mathbf{h} = \frac{1}{2m}(\gamma Q - \rho I)^{-1}\mathbf{y}. \tag{9}$$

is a solution of (5)–(6).

The expression (9) contains the unknown ρ, which we now determine. Let $\mathbf{VDV}^T$ be the spectral decomposition of Q, such that $\mathbf{VDV}^T = Q$, $\mathbf{V}^T\mathbf{V} = \mathbf{VV}^T = I$ and $\mathbf{D}$ is a diagonal matrix with its diagonal elements λ_i being the eigenvalues of Q. Then (7)–(8) can be rewritten as

$$-\mathbf{y}/m + 2\gamma\mathbf{VDV}^T\mathbf{h} - 2\rho\mathbf{VV}^T\mathbf{h} = 0, \tag{10}$$

$$\mathbf{h}^T\mathbf{VV}^T\mathbf{h} - t^2 = 0. \tag{11}$$

Letting $\mathbf{u} = \mathbf{V}^T\mathbf{h}$ and $\mathbf{d} = \mathbf{V}^T\mathbf{y}$, (10)–(11) becomes

$$-\mathbf{d}/m + 2\gamma\mathbf{Du} - 2\rho\mathbf{u} = 0, \tag{12}$$

$$\mathbf{u}^T\mathbf{u} - t^2 = 0. \tag{13}$$

Isolating $\mathbf{u}$ at (12) and substituting it in (13) we get

$$\frac{1}{(2m)^2}\mathbf{d}^T(\gamma\mathbf{D} - \rho I)^{-2}\mathbf{d} - t^2 = 0. \tag{14}$$

Let d_i be the ith component of $\mathbf{d}$. Since the matrix $\mathbf{D}$ is diagonal, (14) is equivalent to

$$\frac{1}{(2m)^2}\sum_{i=1}^{m+u}\frac{d_i^2}{(\gamma\lambda_i - \rho)^2} - t^2 = 0. \tag{15}$$

Equation (15) has multiple ρ solutions. As shown by (Forsythe and Golub 1965, Theorem 2.7), the smallest possible ρ that satisfies (15) also minimizes $\Phi(\mathbf{h}, \rho)$. Thus, our goal is to find the smallest ρ satisfying (15). Since the matrix Q is positive definite, all λ_i's are strictly positive. Therefore, the function

$$f(\rho) = \frac{1}{(2m)^2}\sum_{i=1}^{m+u}\frac{d_i^2}{(\gamma\lambda_i - \rho)^2} - t^2 \tag{16}$$

has the structure depicted in Fig. 3. Considering this structure, our algorithm for finding the smallest ρ such that $f(\rho) = 0$ is as follows: Let $\widetilde{\lambda}$ be the smallest λ_i such that $d_i \neq 0$. We consider the interval $[\widetilde{\lambda} - t_1, \widetilde{\lambda} - t_2]$ such that $t_1 > 0$, $t_2 > 0$, $f(\widetilde{\lambda} - t_1) < 0$, $f(\widetilde{\lambda} - t_2) > 0$ and find the root of f in this domain using a binary search.

[5]Here we assume that the value of ρ satisfying (7)–(8) is not an eigenvalue of γQ and the inverse $(\gamma Q - \rho I)^{-1}$ exists. If this assumption does not hold (this can be easily verified by checking for each eigenvalue of γQ if it satisfies (7)–(8)), then we can slightly perturb the hyperparameter γ to satisfy the assumption.

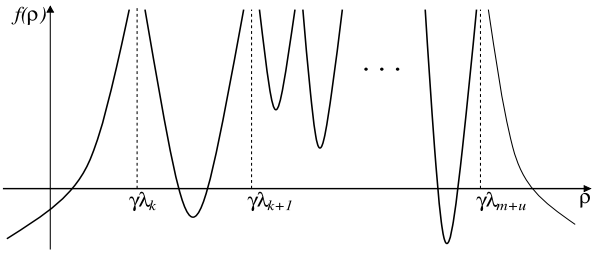

Fig. 3 Structure of the function $f(\rho)$. k is the index of the smallest eigenvalue λ_i such that $d_i \neq 0$

8 A risk bound

In this section we derive a transductive risk bound for the AVR algorithm (5)–(6). Our derivation relies on a known general transductive risk bound for 'unlabeled/labeled (UL) decompositions' of transductive algorithms as discussed in (El-Yaniv and Pechyony 2007).

The soft classification output $\mathbf{h}^*$ of any transductive algorithm can always be represented as $\mathbf{h}^* = K \cdot \boldsymbol{\alpha}$, where K is an $(m+u) \times (m+u)$ matrix depending only on the unlabeled full sample X_{m+u}, and $\boldsymbol{\alpha}$ is an $(m+u) \times 1$ vector that can depend on both S_m and X_u. Such a decomposition is termed a UL (unlabeled-labeled) decomposition (El-Yaniv and Pechyony 2007). Let K_{ij} be the (i, j)th entry of K and $\|K\|_{\text{Fro}}^2 = \sum_{i,j=1}^{m+u} K_{ij}^2$, be the squared Frobenius norm of K. For UL decompositions, the following holds.

Theorem 1 (El-Yaniv and Pechyony 2007) *Let $\mathcal{A}$ be a transductive algorithm and $\mathbf{h}^* = K \cdot \boldsymbol{\alpha}$ be its UL decomposition. Suppose that $\|\boldsymbol{\alpha}\|_2 \leq \mu_1$. Let $c_0 = \sqrt{(32 \ln(4e))/3} < 5.05$ and $W \triangleq 1/m + 1/u$. Let $\widetilde{\mathcal{H}}$ be the set of soft hypotheses that can be generated by the algorithm when operated on any training/test partition. Then,[6] for any $\nu > 0$ and $\delta > 0$, with probability of at least $1 - \delta$ over the choice of the training set of size m from X_{m+u}, for all $\mathbf{h} \in \widetilde{\mathcal{H}}$,*

$$R_u(\mathbf{h}) \leq R_m^{\ell_\nu}(\mathbf{h}) + \frac{\mu_1}{\nu}\sqrt{\frac{2}{mu}\|K\|_{\text{Fro}}^2} + c_0 W\sqrt{\min(m,u)} + \sqrt{2W\ln(1/\delta)}. \tag{17}$$

Since the $(m+u) \times (m+u)$ matrix $\mathbf{V}$ (defined in Sect. 7) consists of orthonormal columns, the solution $\mathbf{h}^*$ of (5)–(6) can be represented as $\mathbf{h}^* = \mathbf{V}\boldsymbol{\alpha}$. The matrix $\mathbf{V}$ depends only on the unlabeled examples. Hence the last equation is a UL decomposition of the AVR algorithm, with $K \triangleq \mathbf{V}$. By (6) we have that $t^2 = \|\mathbf{h}^*\|_2^2 = \boldsymbol{\alpha}^T\mathbf{V}^T\mathbf{V}\boldsymbol{\alpha} = \boldsymbol{\alpha}^T\boldsymbol{\alpha}$. Since each column of $\mathbf{V}$ has unit length, $\|K\|_{\text{Fro}}^2 = m+u$. Substituting $\mu_1 = t$ and $\|K\|_{\text{Fro}}^2 = m+u$ in (17), we obtain[7]

$$R_u(\mathbf{h}) \leq R_m^{\ell_\nu}(\mathbf{h}) + (t/\nu)\sqrt{2W} + c_0 W\sqrt{\min(m,u)} + \sqrt{2W\ln(1/\delta)}. \tag{18}$$

Notice that the matrix Q influences the bound (18) indirectly, through the empirical error term. If t/ν is a constant then the bound (18) converges at rate $1/\sqrt{\min(m,u)}$. In general, there is a trade-off between the values of t and ν. If t is very small then, due to the constraint

[6] The loss function ℓ_ν used in the empirical error term is the hinge loss. For a positive real ν, $\ell_\nu(y_1, y_2) = 0$ if $y_1 y_2 \geq \nu$ and $\ell_\nu(y_1, y_2) = \min\{1,\ 1 - y_1 y_2/\nu\}$.

[7] Using the standard technique of (Bousquet and Elisseeff 2002) (see Theorem 18 there) it is possible to extend (18) to be uniform in ν.

(6), all entries of the hypothesis **h** generated by AVR are very close to zero. In this case, to achieve a small empirical hinge-loss, the value of ν should also be small.

9 Experimental results

We tested the AVR algorithm over 31 binary problems including all 7 datasets from (Chapelle et al. 2006) and all 8 datasets from (Blum and Chawla 2001). We also generated 6 datasets of image classification problems from the COIL-100 dataset (Nene et al. 1996), and took all 10 possible binary problems from the `comp.*` "super-category" in the 20-newsgroups dataset. We randomly subsampled large datasets to contain exactly 1500 examples and in all experiments we used a training set of size 100. We represented text datasets using word-based TF-IDF scores and normalized other datasets using a linear transformation such that the dynamic range of their attributes is [0, 1].

We compared AVR with SVM and TSVM (Collobert et al. 2006).[8] In all problems, with the exception of the text datasets, SVM and TSVM were applied with an RBF kernel. For the text problems, slightly better performance was obtained with a linear kernel. All relevant hyperparameters of the SVM, TSVM and the AVR algorithm were selected using 5-fold cross validation (over the training set), from "reasonable" grids of possible values. For SVM/TSVM, the grid contained 80 possible hyperparameter assignments[9] and for AVR, 72 assignments (as described below).

9.1 On the AVR hyperparameters

The AVR algorithm has a number of parameters. The main parameter, which is left unspecified, is the matrix Q. One natural choice for Q is graph-based using the unnormalized Laplacian.[10] We constructed Q using the unlabeled data X_{m+u} as follows. We computed an $(m+u)\times(m+u)$ similarity matrix S, whose (i,j)th entry is the cosine of the angle between x_i and x_j. Then we built an undirected k-nearest neighbors weighted graph, $G = G(X_{m+u})$, where there is an edge between x_i and x_j iff x_i is among the k most similar (according to S) neighbors of x_j, or vice versa. Edge weights were taken to be the corresponding entries from S.[11] The value of k was selected using 5-fold cross validation from the set $\{5, 10, 100\}$. Let M be the adjacency matrix of G, and let D be the diagonal matrix whose (i,i)th entry is the sum of the ith row of M. Let $L = D - M$ be the unnormalized Laplacian of G.

[8] We applied both SVM and TSVM using the ***UniverSVM*** package of F. Sinz, R. Collobert, J. Weston and L. Bottou, available at http://www.kyb.tuebingen.mpg.de/bs/people/fabee/universvm.html.

[9] The SVM hyperparameters are C (weight of errors of labeled examples) and σ (the "width" of RBF kernel) and we considered all pairs $(C, 1/\sigma^2) \in \{2^{-5}, 2^{-3}, 2^{-1}, 2^1, 2^3, 2^5, 2^7, 2^9\} \times \{2^{-15}, 2^{-13}, 2^{-11}, 2^{-9}, 2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}, 2^1, 2^3\}$. TSVM has additional hyperparameter C^* (weight of errors of unlabeled examples) and we considered all triples $(C, C^*, 1/\sigma^2) = \{2^{-3}, 2^1, 2^5, 2^9\} \times \{2^{-3}, 2^1, 2^5, 2^9\} \times \{2^{-13}, 2^{-9}, 2^{-5}, 2^{-1}, 2^3\}$. For text datasets the hyperparameter σ was not used.

[10] There are, of course, other possibilities for Q, such as a normalized Laplacian, linear models (Wu and Schölkopf 2007) and RBF kernels. We have done preliminary experiments with each one of these possibilities for Q and found that they result in roughly the same performance as the choice of Q based on the unnormalized Laplacian.

[11] Our dataset normalization procedure (described above) implies that the entries of S are non-negative and thus, edge weights in G are non-negative as well.

Remark 1 (On the meaning of the eigenvectors of L) Let

$$\mathcal{G} = \{\mathbf{g} \mid \mathbf{g} \in \mathbb{R}^{m+u}, \|\mathbf{g}\|_2 = 1\}.$$

For any $\mathbf{g} \in \mathcal{G}$, let $\mathbf{g}^T L\mathbf{g} = \sum_{i,j=1}^{m+u}(g_i - g_j)^2 m_{ij}$ be its "soft smoothness" w.r.t. the graph $G(X_{m+u})$, where m_{ij} is the (i, j)th entry of M. By the Rayleigh-Ritz theorem (Horn and Johnson 1990), the smallest eigenvalue of L is $\lambda_1 = \min_{\mathbf{g}\in\mathcal{G}} \mathbf{g}^T L\mathbf{g}$ and its eigenvector is $\mathbf{v}_1 = \arg\min_{\mathbf{g}\in\mathcal{G}} \mathbf{g}^T L\mathbf{g}$. Thus, the first eigenvector $\mathbf{v}_1$ has the maximal smoothness (of value λ_1). A generalization of this theorem in (Horn and Johnson 1990) implies that for any $1 \leq r \leq m+u$, $\lambda_r = \min_{\mathbf{g}\in\mathcal{G}, \mathbf{g}^T\mathbf{v}_1=0,\ldots,\mathbf{g}^T\mathbf{v}_{r-1}=0} \mathbf{g}^T L\mathbf{g}$, and the minimum is achieved by $\mathbf{v}_r$. Thus, the r-th largest eigenvector is the maximally smooth vector (with smoothness λ_r) among all the vectors that are orthogonal to the $r-1$ maximally smooth vectors. By taking $Q \triangleq L$, we therefore prioritize highly smooth equivalent classes.

Although we would like to assign $Q \triangleq L$, it is well known that L is positive *semi*-definite. We need Q to be positive-definite to ensure a finite volume.[12] To this end, we truncate the larger eigenvectors of L using the following simple heuristic. We fix two parameters: a threshold $\tau > 0$ and $0 < c < 1$, and truncate the $l = \lceil c \cdot (m+u) \rceil$ largest eigenvectors to length τ. Let μ be the "stretch factor" of the lth largest vector (new length/old length). To preserve length proportions among the $m+u-l$ smallest eigenvectors we stretch (or shrink) them by a factor μ. In our experiments we selected the values of the hyperparameters τ and c from the sets $\{1, 10\}$ and $\{0.05, 0.1, 0.9, 0.95\}$, respectively, and set $t = 1$. Finally, the hyperparameter γ was selected from $\{0.01/m, 1/m, 100/m\}$.

Remark 2 (On the computational complexity of AVR) By our construction of Q the computational complexity of AVR is dominated by the eigendecomposition of the graph Laplacian[13] L. In general the complexity of this decomposition is $O((m+u)^3)$. For small $k \ll m+u$, the matrix L is very sparse and the eigendecomposition can be computed faster. In Sect. 10 we discuss a fast method for constructing Q without performing costly eigendecompositions.

Remark 3 (On the choice of the loss function) The solution (9) of the AVR optimization problem involves the inversion of the $(m+u) \times (m+u)$ matrix $\gamma Q - \rho I$. This operation is computationally expensive and has time complexity of $O((m+u)^{2.376})$ (Coppersmith and Winograd 1990). Let $\{(\lambda_i, \mathbf{v}_i)\}_{i=1}^{m+u}$ be the eigendecomposition of Q. Recall that by our method of constructing Q, we know its eigendecomposition before computing $(\gamma Q - \rho I)^{-1}$. Since

$$(\gamma Q - \rho I)^{-1} = \sum_{i=1}^{m+u} \frac{1}{\gamma\lambda_i - \rho} \mathbf{v}_i \mathbf{v}_i^T, \tag{19}$$

given the eigendecomposition of Q, the inverse $(\gamma Q - \rho I)^{-1}$ can be computed fast. Note that the eigendecomposition of Q is independent of the training/test partition and the choice of the hyperparameters γ, τ and c. Thus, we can reuse the eigendecomposition of Q for different values of γ, τ, c and different training/test partitions and speed-up our experiments.

[12] If Q is semi-definite then its smallest eigenvalue λ_1 is zero. In this case the length $(1/\sqrt{\lambda_1})$ of the principal vector $\mathbf{v}_1$ (corresponding to λ_1) of the ellipse $\mathcal{E}(\mathcal{H}_Q)$ (see Fig. 2) is infinite.

[13] Recall that this eigendecomposition is required in order to make L positive definite.

This reuse would be impossible if instead of the linear loss $-y_i h_i$ we took the commonly used squared loss[14] $(y_i - h_i)^2$, resulting in an order of magnitude slow-down.

9.2 Results

Our results for the 31 datasets appear in Tables 1 and 2. Each experiment was performed 12 times with different random train/test partitions. In Tables 1 and 2, each entry is an average ($\pm$ standard error of the mean) of these 12 experiments. It is evident that AVR overwhelmingly outperforms SVM and TSVM on the dataset collection of Table 1. In particular, AVR exhibits excellent performance in text categorization and image classification. However, AVR is significantly inferior to SVM/TSVM on the UCI datasets of Table 2.

Table 1 Results for three dataset collections

Dataset	SVM	TSVM	AVR
Datasets from (Chapelle et al. 2006)			
g241c	23.07 ± 0.44	$\mathbf{18.62 \pm 1.09}$	$\mathbf{20.21 \pm 1.40}$
g241n	25.26 ± 0.40	23.29 ± 0.87	$\mathbf{12.1 \pm 0.87}$
digit	6.03 ± 0.53	$\mathbf{5.18 \pm 0.73}$	$\mathbf{4.21 \pm 0.56}$
usps	9.07 ± 0.45	8.11 ± 0.95	$\mathbf{6.23 \pm 0.46}$
coil	17.41 ± 1.31	17.15 ± 1.39	$\mathbf{5.89 \pm 0.42}$
bci	$\mathbf{31.75 \pm 1.21}$	$\mathbf{32.92 \pm 0.47}$	48.94 ± 0.99
text	$\mathbf{25.47 \pm 0.74}$	$\mathbf{24.05 \pm 0.88}$	$\mathbf{24.73 \pm 0.47}$
Image datasets			
coil1	$\mathbf{12.35 \pm 0.45}$	$\mathbf{12.21 \pm 0.39}$	$\mathbf{11.63 \pm 0.37}$
coil2	9.37 ± 0.23	8.25 ± 0.34	$\mathbf{2.07 \pm 0.52}$
coil3	20.04 ± 0.56	18.44 ± 0.61	$\mathbf{12.17 \pm 0.71}$
coil4	12.35 ± 0.67	9.73 ± 0.35	$\mathbf{5.46 \pm 0.56}$
coil5	25.75 ± 1.46	24.69 ± 1.89	$\mathbf{15.62 \pm 0.87}$
coil6	24.46 ± 1.07	23.13 ± 0.90	$\mathbf{8.50 \pm 1.39}$
Text (20 Newsgroups)			
graphics/misc	19.79 ± 1.46	17.54 ± 1.09	$\mathbf{14.76 \pm 0.34}$
graphics/pc	16.86 ± 1.79	13.96 ± 1.39	$\mathbf{9.55 \pm 0.30}$
graphics/mac	12.85 ± 1.68	10.39 ± 1.28	$\mathbf{7.64 \pm 0.54}$
graphics/X	20.99 ± 2.12	16.42 ± 1.17	$\mathbf{14.36 \pm 0.76}$
misc/pc	21.25 ± 1.79	19.40 ± 1.30	$\mathbf{16.12 \pm 0.68}$
misc/mac	13.74 ± 1.70	$\mathbf{11.83 \pm 1.24}$	$\mathbf{10.90 \pm 0.35}$
misc/X	16.91 ± 2.13	$\mathbf{13.63 \pm 1.44}$	$\mathbf{12.85 \pm 0.49}$
pc/mac	$\mathbf{23.41 \pm 2.00}$	$\mathbf{20.40 \pm 1.21}$	$\mathbf{20.42 \pm 0.78}$
pc/X	9.79 ± 2.27	8.76 ± 1.38	$\mathbf{5.74 \pm 0.21}$
mac/X	10.73 ± 2.55	8.27 ± 1.35	$\mathbf{4.28 \pm 0.28}$

[14] If we use the squared loss, then instead of (9) we would obtain $\mathbf{h} = \frac{1}{2m}(\gamma Q - \rho I + I^*)^{-1}\mathbf{y}$, where I^* is a diagonal matrix whose (i,i)th entry equals 1, if the ith example is in the training set, and zero otherwise. The inverse in the last expression cannot be computed using the eigendecomposition of Q.

Table 2 UCI datasets taken from (Blum and Chawla 2001)

Dataset	SVM	TSVM	AVR
pima	**27.96 ± 1.06**	**27.97 ± 1.07**	36.78 ± 0.83
bupa	**34.52 ± 0.86**	**32.99 ± 1.09**	36.63 ± 1.20
mush	**3.26 ± 0.41**	**2.88 ± 0.47**	**2.85 ± 0.49**
musk	**12.44 ± 0.70**	**11.75 ± 0.59**	15.62 ± 0.78
monk	**0.58 ± 0.36**	1.93 ± 0.82	20.16 ± 0.26
ionosphere	**7.93 ± 0.66**	**7.44 ± 0.75**	16.50 ± 0.61
tae	**26.96 ± 1.73**	**29.08 ± 1.67**	37.75 ± 2.05
voting	**5.26 ± 0.40**	**4.64 ± 0.26**	7.06 ± 0.43

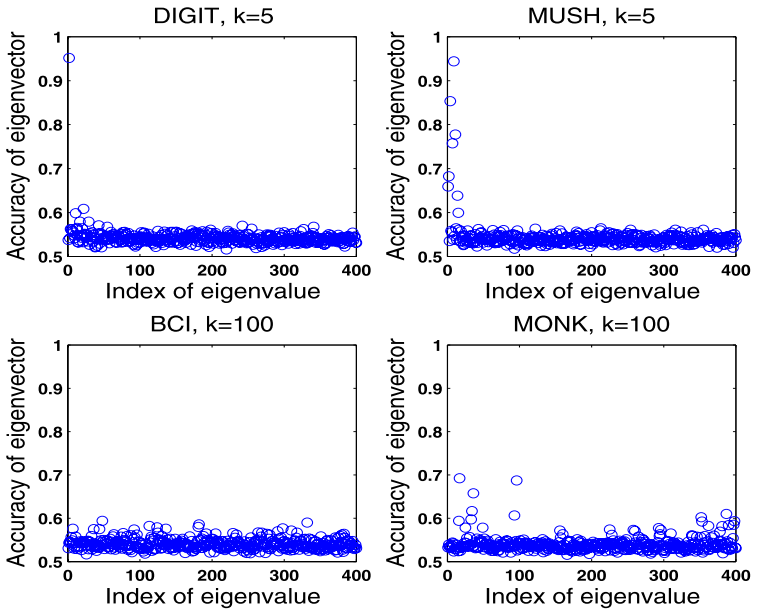

Fig. 4 Accuracy of the eigenvectors of Q

9.3 Analysis of results

We investigated further cases where the AVR succeeded and failed and found two empirical characterizations of its performance.

9.3.1 Accuracy of eigenvectors

Let $\{\lambda_i, \mathbf{v}_i\}_{i=1}^{m+u}$ be the eigenvectors and eigenvalues of Q, such that $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{m+u}$. Since $\mathbf{v}_i \in \mathbb{R}^{m+u}$, we consider it as a vector of soft classifications of the full sample examples from X_{m+u}. In particular, we consider the jth entry of $\mathbf{v}_i$ as the soft classification of x_j. For each $1 \leq i \leq m+u$ we computed the training accuracy of $\mathbf{v}_i$ and $-\mathbf{v}_i$ and took the best accuracy among these two numbers. The resulting graphs of eigenvector training accuracies, averaged over 12 training/test partitions, are shown in Fig. 4 for 4 datasets. For each

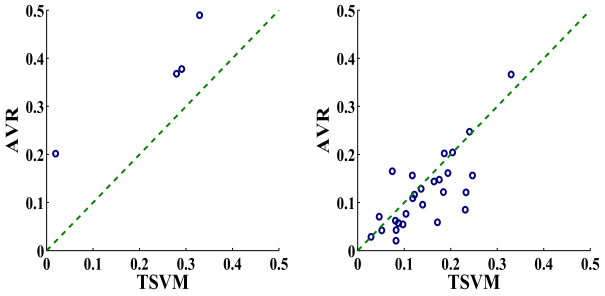

Fig. 5 Comparison of AVR versus TSVM: (**a**) TSVM looses to SVM; (**b**) TSVM wins over SVM

dataset we chose the value of k yielding the best performance in hindsight. We truncated the graphs to include the 400 smallest eigenvalues, since the accuracies of the eigenvectors corresponding to larger eigenvalues are almost always as those obtained by the eigenvectors corresponding to the eigenvalues with indices 200–400.

Figure 4 shows that in two datasets where AVR succeeds (DIGIT and MUSH) there are a few very accurate eigenvectors, which correspond to small eigenvalues of Q. Moreover, in these datasets there is a large gap in the accuracy of these eigenvectors and the others and there are no accurate eigenvectors corresponding to large eigenvalues. In contrast, in datasets where AVR failed (BCI and MONK) the accuracy of the eigenvectors corresponding to small eigenvalues is quite low. Qualitatively similar effects were observed in all the other datasets.

The above characterization in terms of the accuracies of the eigenvectors of Q suggests the following heuristic to quickly assess whether we should use AVR or large margin methods (SVM or TSVM). If the accuracy of the leading eigenvectors of Q is high relative to the accuracy of the large-margin methods, then run AVR with cross validation to determine its best parameters. Otherwise, use a large-margin method.

9.3.2 Magnification of TSVM success and failure

Using the results of Tables 1 and 2 we divide all 31 datasets into two categories. The first category consists of the datasets where SVM outperformed TSVM. The second category consists of those where TSVM outperformed SVM. There are 4 datasets in the former category and 27 in the latter. Note that in this partition we measure performance by considering only average errors and ignore standards error of the means.

A comparison of AVR and TSVM over datasets of these two categories is depicted in Fig. 5 using scatter plots. In these plots each point represents a comparison on a single dataset. If the point falls below the dashed line then AVR outperformed TSVM, and vice versa. It is evident that if SVM outperformed TSVM then TSVM also outperformed AVR. Conversely, if TSVM outperformed SVM, then in the significant majority of the datasets, AVR also outperformed TSVM. Thus, in cases where transductive learning was effective (in the sense that TSVM outperformed SVM), the AVR algorithm magnified the success of TSVM, and vice versa,

10 Concluding remarks

We developed a new transductive technique based on a large volume principle. The new technique is well motivated using the transductive maximal power inference. The resulting AVR algorithm that approximates this scheme is extremely successful in three (out of the

four) sets of problems we examined (in particular, in text categorization and image classification problems) and fails in a set of UCI problems. The main questions are: why does AVR fail in the last set? How can we make a better data-dependent selection of the ellipsoid matrix Q?

One possible direction could be to explicitly design Q matrices by encoding in eigenvectors and eigenvalues useful prior knowledge and information about the given data. We note that such constructions can also be beneficial from a computational complexity viewpoint since they would save the need for the spectral decompositions we currently perform.

It would be very interesting to identify datasets' characteristics that give an advantage to either the large margin or the large volume principle. In this regard, we note that in our comparison here these two principles were applied on different hypothesis spaces, namely, (kernelized) hyperplanes in the case of large margin and arbitrary soft response vectors in the case of large volume (which has much larger capacity). It would be of interest to compare these two principles w.r.t. the same space. We observed that the AVR algorithm magnified the success and failure of TSVM. We plan to further investigate this interesting effect. further. Finally, a technical interesting question is how to better approximate or provide approximation guarantees for volume assessments in the context of our elliptic hypothesis classes.

References

Bennett, C., & Bredensteiner, E. (2000). Geometry in learning. In *Geometry at work* (pp. 132–145). Washington: Mathematical Association of America.

Blum, A., & Chawla, S. (2001). Learning from labeled and unlabeled data using graph mincuts. In *ICML* (pp. 19–26).

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, *2*, 499–526.

Chapelle, O., Schölkopf, B., & Zien, A. (2006). *Semi-supervised learning*. Cambridge: MIT Press.

Coppersmith, D., & Winograd, S. (1990). Matrix multiplication via arithmetic progressions. *Journal of Symbolic Computation*, *9*, 251–280.

Derbeko, P., El-Yaniv, R., & Meir, R. (2004). Explicit learning curves for transduction and application to clustering and compression algorithms. *Journal of Artificial Intelligence Research*, *22*, 117–142.

El-Yaniv, R., & Pechyony, D. (2007). Transductive Rademacher complexity and its applications. In *COLT*.

Forsythe, G., & Golub, G. (1965). On the stationary values of a second-degree polynomial on the unit sphere. *Journal of the Society for Industrial and Applied Mathematics*, *13*(4), 1050–1068.

Gander, W., Golub, G., & von Matt, U. (1989). A constrained eigenvalue problem. *Linear Algebra and Its Applications*, *114/115*, 815–839.

Graepel, T., Herbrich, R., & Obermayer, K. (1999). Bayesian transduction. In *NIPS* (pp. 456–462).

Hanneke, S. (2006). An analysis of graph cut size for transductive learning. In *ICML*.

Horn, R., & Johnson, C. (1990). *Matrix analysis*. Cambridge: Cambridge University Press.

Lovasz, L., & Vempala, S. (2006). Simulated annealing in convex bodies and an $O^*(n^4)$ volume algorithm. *Journal of Computer and System Sciences*, *72*(2), 392–417.

Nene, S., Nayar, S., & Murase, H. (1996). *Columbia object image library (coil-100)* (Technical Report CUCS-006-96). Columbia University.

Collobert, F., Sinz, R., Weston, J., Bottou, L., (2006). Large scale transductive svms. *Journal of Machine Learning Research*, *7*, 1687–1712.

Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, *2*, 45–66.

Vapnik, V. N. (1982). *Estimation of dependences based on empirical data*. New York: Springer.

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley Interscience.

Weston, J., Collobert, R., Sinz, F., Bottou, L., & Vapnik, V. (2006). On the inference with universum. In *ICML*.

Wu, M., & Schölkopf, B. (2007). Transductive classification with local learning regularization. In *AISTATS*.