Ranking and combining multiple predictors without labeled data

Fabio Parisi[a,1], Francesco Strino[a,1], Boaz Nadler[b], and Yuval Kluger[a,c,2]

[a]Department of Pathology, Yale University School of Medicine, New Haven, CT 06520; [b]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel; and [c]New York University Center for Health Informatics and Bioinformatics, New York University Langone Medical Center, New York, NY 10016

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved December 17, 2013 (received for review November 1, 2012)

In a broad range of classification and decision-making problems, one is given the advice or predictions of several classifiers, of unknown reliability, over multiple questions or queries. This scenario is different from the standard supervised setting, where each classifier's accuracy can be assessed using available labeled data, and raises two questions: Given only the predictions of several classifiers over a large set of unlabeled test data, is it possible to (*i*) reliably rank them and (*ii*) construct a metaclassifier more accurate than most classifiers in the ensemble? Here we present a spectral approach to address these questions. First, assuming conditional independence between classifiers, we show that the off-diagonal entries of their covariance matrix correspond to a rank-one matrix. Moreover, the classifiers can be ranked using the leading eigenvector of this covariance matrix, because its entries are proportional to their balanced accuracies. Second, via a linear approximation to the maximum likelihood estimator, we derive the Spectral Meta-Learner (SML), an unsupervised ensemble classifier whose weights are equal to these eigenvector entries. On both simulated and real data, SML typically achieves a higher accuracy than most classifiers in the ensemble and can provide a better starting point than majority voting for estimating the maximum likelihood solution. Furthermore, SML is robust to the presence of small malicious groups of classifiers designed to veer the ensemble prediction away from the (unknown) ground truth.

spectral analysis | classifier balanced accuracy | unsupervised learning | cartels | crowdsourcing

Every day, multiple decisions are made based on input and suggestions from several sources, either algorithms or advisers, of unknown reliability. Investment companies handle their portfolios by combining reports from several analysts, each providing recommendations on buying, selling, or holding multiple stocks (1, 2). Central banks combine surveys of several professional forecasters to monitor rates of inflation, real gross domestic product growth, and unemployment (3–6). Biologists study the genomic binding locations of proteins by combining or ranking the predictions of several peak detection algorithms applied to large-scale genomics data (7). Physician tumor boards convene a number of experts from different disciplines to discuss patients whose diseases pose diagnostic and therapeutic challenges (8). Peer-review panels discuss multiple grant applications and make recommendations to fund or reject them (9). The examples above describe scenarios in which several human advisers or algorithms provide their predictions or answers to a list of queries or questions. A key challenge is to improve decision making by combining these multiple predictions of unknown reliability. Automating this process of combining multiple predictors is an active field of research in decision science (cci.mit.edu/research), medicine (10), business (refs. 11 and 12 and www.kaggle.com/competitions), and government (www.iarpa.gov/Programs/ia/ACE/ace.html and www.goodjudgmentproject.com), as well as in statistics and machine learning.

Such scenarios, whereby advisers of unknown reliability provide potentially conflicting opinions, or propose to take opposite actions, raise several interesting questions. How should the decision maker proceed to identify who, among the advisers, is the most reliable? Moreover, is it possible for the decision maker to cleverly combine the collection of answers from all of the advisers and provide even more accurate answers?

In statistical terms, the first question corresponds to the problem of estimating prediction performances of preconstructed classifiers (e.g., the advisers) in the absence of class labels. Namely, each classifier was constructed independently on a potentially different training dataset (e.g., each adviser trained on his/her own using possibly different sources of information), yet they are all being applied to the same new test data (e.g., list of queries) for which labels are not available, either because they are expensive to obtain or because they will only be available in the future, after the decision has been made. In addition, the accuracy of each classifier on its own training data is unknown. This scenario is markedly different from the standard supervised setting in machine learning and statistics. There, classifiers are typically trained on the same labeled data and can be ranked, for example, by comparing their empirical accuracy on a common labeled validation set. In this paper we show that under standard assumptions of independence between classifier errors their unknown performances can still be ranked even in the absence of labeled data.

The second question raised above corresponds to the problem of combining predictions of preconstructed classifiers to form a metaclassifier with improved prediction performance. This problem arises in many fields, including combination of forecasts in decision science and crowdsourcing in machine learning, which have each derived different approaches to address it. If we had external knowledge or historical data to assess the reliability of the available classifiers we could use well-established solutions relying on panels of experts or forecast combinations (11–14). In

Significance

A key challenge in a broad range of decision-making and classification problems is how to rank and combine the possibly conflicting suggestions of several advisers of unknown reliability. We provide mathematical insights of striking conceptual simplicity that explain mutual relationships between independent advisers. These insights enable the design of efficient, robust, and reliable methods to rank the advisers' performances and construct improved predictions in the absence of ground truth. Furthermore, these methods are robust to the presence of small subgroups of malicious advisers (cartels) attempting to veer the combined decisions to their interest.

Author contributions: F.P., F.S., B.N., and Y.K. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]F.P. and F.S. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: yuval.kluger@yale.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1219097111/-/DCSupplemental.

our problem such knowledge is not always available and thus these solutions are in general not applicable. The oldest solution that does not require additional information is majority voting, whereby the predicted class label is determined by a rule of majority, with all advisers assigned the same weight. More recently, iterative likelihood maximization procedures, pioneered by Dawid and Skene (15), have been proposed, in particular in crowdsourcing applications (16–23). Owing to the nonconvexity of the likelihood function, these techniques often converge only to a local, rather than global, maximum and require careful initialization. Furthermore, there are typically no guarantees on the quality of the resulting solution.

In this paper we address these questions via a spectral analysis that yields four major insights:

1. Under standard assumptions of independence between classifier errors, in the limit of an infinite test set, the off-diagonal entries of the population covariance matrix of the classifiers correspond to a rank-one matrix.
2. The entries of the leading eigenvector of this rank-one matrix are proportional to the balanced accuracies of the classifiers. Thus, a spectral decomposition of this rank-one matrix provides a computationally efficient approach to rank the performances of an ensemble of classifiers.
3. A linear approximation of the maximum likelihood estimator yields an ensemble learner whose weights are proportional to the entries of this eigenvector. This represents an efficient, easily constructed, unsupervised ensemble learner, which we term Spectral Meta-Learner (SML).
4. An interest group of conspiring classifiers (a cartel) that maliciously attempts to veer the overall ensemble solution away from the (unknown) ground truth leads to a rank-two covariance matrix. Furthermore, in contrast to majority voting, SML is robust to the presence of a small-enough cartel whose members are unknown.

In addition, we demonstrate the advantages of spectral approaches based on these insights, using both simulated and real-world datasets. When the independence assumptions hold approximately, SML is typically better than most classifiers in the ensemble and their majority vote, achieving results comparable to the maximum likelihood estimator (MLE). Empirically, we find SML to be a better starting point for computing the MLE that consistently leads to improved performance. Finally, spectral approaches are also robust to cartels and therefore helpful in analyzing surveys where a biased subgroup of advisers (a cartel) may have corrupted the data.

Problem Setup

For simplicity, we consider the case of questions with yes/no answers. Hence, the advisers, or algorithms, provide to each query only one of two possible answers, either +1 (positive) or −1 (negative). Following standard statistical terminology, the advisers or algorithms are called "binary classifiers," and their answers are termed "predicted class labels." Each question is represented by a feature vector x contained in a feature space $\mathcal{X}$.

In detail, let $\{f_i\}_{i=1}^{M}$ be M binary classifiers of unknown reliability, each providing predicted class labels $f_i(x_k)$ to a set of S instances $D=\{x_k\}_{k=1}^{S}\subset\mathcal{X}$, whose vector of true (unknown) class labels is denoted by $\mathbf{y}=(y_1,\ldots,y_S)$.We assume that each classifier $f_i:\mathcal{X}\to\{-1,1\}$ was trained in a manner undisclosed to us using its own labeled training set, which is also unavailable to us. Thus, we view each classifier as a black-box function of unknown classification accuracy.

Using only the predictions of the M binary classifiers on the unlabeled set D and without access to any labeled data, we consider the two problems stated in the introduction: (i) Rank the performances of the M classifiers and (ii) combine their predictions to provide an improved estimate $\hat{\mathbf{y}}=(\hat{y}_1,\ldots,\hat{y}_S)$ of the true class label vector $\mathbf{y}$.

We represent an instance and class label pair $(X,Y)\in\mathcal{X}\times\{-1,1\}$ as a random vector with probability density function $p(x,y)$, and with marginals $p_X(x)$ and $p_Y(y)$.

In the present study, we measure the performance of a binary classifier f by its balanced accuracy π, defined as

$$\pi=\frac{\text{sensitivity}+\text{specificity}}{2}=\frac{1}{2}(\psi+\eta), \qquad [1]$$

where ψ and η are its sensitivity (fraction of correctly predicted positives) and specificity (fraction of correctly predicted negatives). Formally, these quantities are defined as

$$\psi=\Pr[f(X)=Y|Y=1], \quad \text{and} \quad \eta=\Pr[f(X)=Y|Y=-1]. \qquad [2]$$

Assumptions. In our analysis we make the following two assumptions: (i) The S unlabeled instances $x_k\in D$ are independent and identically distributed realizations from the marginal distribution $p_X(x)$ and (ii) the M classifiers are conditionally independent, in the sense that prediction errors made by one classifier are independent of those made by any other classifier. Namely, for all $1\le i\ne j\le M$, and for each of the two class labels, with $a_i,a_j\in\{-1,1\}$

$$\Pr\left[f_i(X)=a_i,f_j(X)=a_j\middle|Y\right]=\Pr[f_i(X)=a_i|Y]\cdot\Pr\left[f_j(X)=a_j\middle|Y\right]. \qquad [3]$$

Classifiers that are nearly conditionally independent may arise, for example, from advisers who did not communicate with each other, or from algorithms that are based on different design principles or independent sources of information. Note that these assumptions appear also in other works considering a setting similar to ours (15, 23), as well as in supervised learning, the development of classifiers (e.g., Naïve Bayes), and ensemble methods (24).

Ranking of Classifiers

To rank the M classifiers without any labeled data, in this paper we present a spectral approach based on the covariance matrix of the M classifiers. To motivate our approach it is instructive to first study its asymptotic structure as the number of unlabeled test data tends to infinity, $|D|=S\to\infty$. Let Q be the $M\times M$ population covariance matrix of the M classifiers, whose entries are defined as

$$q_{ij}=\mathbb{E}\left[(f_i(X)-\mu_i)\left(f_j(X)-\mu_j\right)\right], \qquad [4]$$

where $\mathbb{E}$ denotes expectation with respect to the density $p(x,y)$ and $\mu_i=\mathbb{E}[f_i(X)]$.

The following lemma, proven in *SI Appendix*, characterizes the relation between the matrix Q and the balanced accuracies of the M classifiers:

Lemma 1. *The entries q_{ij} of Q are equal to*

$$q_{ij}=\begin{cases} 1-\mu_i^2 & i=j \\ (2\pi_i-1)\left(2\pi_j-1\right)\left(1-b^2\right) & \text{otherwise} \end{cases} \qquad [5]$$

where $b\in(-1,1)$ is the class imbalance,

$$b=\Pr[Y=1]-\Pr[Y=-1]. \qquad [6]$$

The key insight from this lemma is that the off-diagonal entries of Q are identical to those of a rank-one matrix $R=\lambda\mathbf{v}\mathbf{v}^T$ with unit-norm eigenvector $\mathbf{v}$ and eigenvalue

$$\lambda = \left(1 - b^2\right) \cdot \sum_{i=1}^{M} (2\pi_i - 1)^2. \qquad [7]$$

Importantly, up to a sign ambiguity, the entries of $\mathbf{v}$ are proportional to the balanced accuracies of the M classifiers,

$$v_i \propto (2\pi_i - 1). \qquad [8]$$

Hence, the M classifiers can be ranked according to their balanced accuracies by sorting the entries of the eigenvector $\mathbf{v}$.

Although typically neither Q nor $\mathbf{v}$ is known, both can be estimated from the finite unlabeled dataset D. We denote the corresponding sample covariance matrix by $\hat{Q}$. Its entries are

$$\hat{q}_{ij} = \frac{1}{S-1} \sum_{k=1}^{S} \left(f_i(x_k) - \hat{\mu}_i\right)\left(f_j(x_k) - \hat{\mu}_j\right),$$

where $\hat{\mu}_i = \frac{1}{S}\sum_k f_i(x_k)$. Under our assumptions, $\hat{Q}$ is an unbiased estimate of Q, e.g., $\mathbb{E}[\hat{Q}] = Q$. Moreover, the variances of its off-diagonal entries are given by

$$Var\left[\hat{q}_{ij}\right] = \frac{\left(1 - \mu_i^2\right) \cdot \left(1 - \mu_j^2\right)}{S-1} + \frac{q_{ij}}{S}\left(4\mu_i\mu_j - \frac{S-2}{S-1} q_{ij}\right). \qquad [9]$$

In particular, $\hat{q}_{ij} - q_{ij} = O(1/\sqrt{S})$ and asymptotically $\hat{Q} \to Q$ as $S \to \infty$. Hence, for a sufficiently large unlabeled set D, it should be possible to accurately estimate from $\hat{Q}$ the eigenvector $\mathbf{v}$ and consequently the ranking of the M classifiers.

One possible approach is to construct an estimate $\hat{R}$ of the rank-one matrix R and then compute its leading eigenvector. Given that $\mathbb{E}[\hat{Q}] = Q$, for all $i \neq j$ we may estimate $\hat{r}_{ij} = \hat{q}_{ij}$, and we only need to estimate the diagonal entries of R. A computationally efficient way to do this, by solving a set of linear equations, is based on the following observation: Upon the change of variables $\left|r_{ij}\right| = e^{t_i} \cdot e^{t_j}$, we have for all $i \neq j$,

$$\log\left|r_{ij}\right| - t_i - t_j = \log\left|q_{ij}\right| - t_i - t_j = 0.$$

Hence, if we knew q_{ij} we could find the vector $\mathbf{t}$ by solving the above system of equations. In practice, because we only have access to $\hat{q}_{ij}$ we thus look for a vector $\hat{\mathbf{t}}$ with small residual error in the above $M(M-1)/2$ equations. We then estimate the diagonal entries by $\hat{r}_{ii} = \exp(2\hat{t}_i)$ and proceed with eigendecomposition of $\hat{R}$. Further details on this and other approaches to estimate $\mathbf{v}$ appear in *SI Appendix*.

Next, let us briefly discuss the error in this approach. First, because $\hat{Q} \to Q$ as $S \to \infty$, it follows that $\hat{\mathbf{t}} \to \mathbf{t}$ and consequently $\hat{R} \to R$. Hence, asymptotically we perfectly recover the correct ranking of the M classifiers. Because R is rank-one, $\hat{R} - R = O\left(1/\sqrt{S}\right)$ and both R and $\hat{R}$ are symmetric, as shown in *SI Appendix*, the leading eigenvector is stable to small perturbations. In particular, $\hat{\mathbf{v}} - \mathbf{v} = O\left(\frac{1}{\lambda}\frac{1}{\sqrt{S}}\right)$. Finally, note that if all classifiers are better than random and the class imbalance is bounded away from ±1, then we have a large spectral gap with $\lambda = O(M)$.

SML

Next, we turn to the problem of constructing a metalearner expected to be more accurate than most (if not all) of the M classifiers in the ensemble. In our setting, this is equivalent to estimating the S unknown labels $y_1, \ldots, y_S$ by combining the labels predicted by the M classifiers.

The standard approach to this task is to determine the MLE $\hat{\mathbf{y}}^{(\mathrm{ML})}$ of the true class labels $\mathbf{y}$ for all of the unlabeled instances (15). Under the assumption of independence between classifier errors and between instances, the overall likelihood is the product of the likelihoods of the S individual instances, where the likelihood of a label y for an instance x is

$$\mathfrak{L}(f_1(x), \ldots, f_M(x); y) = \prod_{i=1}^{M} \Pr(f_i(x)|y). \qquad [10]$$

As shown in *SI Appendix*, the MLE can be written as a weighted sum of the binary labels $f_i(x) \in \{-1, 1\}$, with weights that depend on the sensitivities ψ_i and specificities η_i of the classifiers. For an instance x,

$$\begin{aligned} \hat{y}^{(\mathrm{ML})} &= \underset{y}{\operatorname{argmax}}\, \mathfrak{L}(f_1(x), \ldots, f_M(x); y) \\ &= \operatorname{sign}\left(\sum_{i=1}^{M} f_i(x) \log\alpha_i + \log\beta_i\right), \end{aligned} \qquad [11]$$

where

$$\alpha_i = \frac{\psi_i \eta_i}{(1-\psi_i)(1-\eta_i)}, \quad \beta_i = \frac{\psi_i(1-\psi_i)}{\eta_i(1-\eta_i)}. \qquad [12]$$

Eq. **11** shows that the MLE is a linear ensemble classifier, whose weights depend, unfortunately, on the unknown specificities and sensitivities of the M classifiers.

The common approach, pioneered by Dawid and Skene (15), is to look for all S labels and M classifier specificities and sensitivities that jointly maximize the likelihood. Given an estimate of the true class labels, it is straightforward to estimate each classifier sensitivity and specificity. Similarly, given estimates of ψ_i and η_i, the corresponding estimates of $\mathbf{y}$ are easily found via Eq. **11**. Hence, the MLE is typically approximated by expectation maximization (EM) (18–21, 23).

As is well known, the EM procedure is guaranteed to increase the likelihood at each iteration till convergence. However, its key limitation is that owing to the nonconvexity of the likelihood function the EM iterations often converge to a local (rather than global) maximum.

Importantly, the EM procedure requires an initial guess of the true labels $\mathbf{y}$. A common choice is the simple majority rule of all classifiers. As noted in previous studies, majority voting may be suboptimal, and starting from it, the EM procedure may converge to suboptimal local maxima (23). Thus, it is desirable, and sometimes crucial, to initialize the EM algorithm with an estimate $\hat{\mathbf{y}}$ that is close to the true label $\mathbf{y}$.

Using the eigenvector described in the previous section, we now construct an initial guess that is typically more accurate than majority voting. To this end, note that a Taylor expansion of the unknown coefficients α_i and β_i in Eq. **12** around $(\psi_i, \eta_i) = (1/2, 1/2)$ gives, up to second-order terms, $O((\psi_i - 1/2)^2, (\eta_i - 1/2)^2, (\psi_i - \ 1/2) \cdot (\eta_i - 1/2))$,

$$\alpha_i \approx 1 + 4(\psi_i + \eta_i - 1) = 1 + 4(2\pi_i - 1), \qquad \beta_i \approx 1. \qquad [13]$$

Hence, combining Eq. **13** with a first-order Taylor expansion of the argument inside the sign function in Eq. **11**, around $(\psi_i, \eta_i) = (1/2, 1/2)$ yields

$$\hat{y}_k^{(\mathrm{ML})} \approx \operatorname{sign}\left(\sum_{i=1}^{M} f_i(x_k)(2\pi_i - 1)\right). \qquad [14]$$

Recall that by *Lemma 1* up to a sign ambiguity the entries of the leading eigenvector of R are proportional to the balanced accuracies of the classifiers, $v_i \propto (2\pi_i - 1)$. This sign ambiguity can be easily resolved if we assume, for example, that most classifiers are better than random. Replacing $2\pi_i - 1$ in Eq. **14** by the eigenvector entries $\hat{v}_i$ of an estimate of R yields a novel spectral-based ensemble classifier, which we term the SML,

$$\hat{y}_k^{(\mathrm{SML})} = \mathrm{sign}\left(\sum_{i=1}^{M} f_i(x_k)\cdot \hat{v}_i\right). \quad \textbf{[15]}$$

Intuitively, we expect SML to be more accurate than majority voting as it attempts to give more weight to more accurate classifiers. *Lemma S2* in *SI Appendix* provides insights on the improved performance achieved by SML in the special case when all algorithms but one have the same sensitivity and specificity. Numerical results for more general cases are described in *Simulations*, where we also show that empirically, on several real data problems, SML provides a better initial guess than majority voting for EM procedures that iteratively estimate the MLE.

Learning in the Presence of a Malicious Cartel

Consider a scenario whereby a small fraction r of the M classifiers belong to a conspiring cartel (e.g., representing a junta or an interest group), maliciously designed to veer the ensemble solution toward the cartel's target and away from the truth. The possibility of such a scenario raises the following question: How sensitive are SML and majority voting to the presence of a cartel? In other words, to what extent can these methods ignore, or at least substantially reduce, the effect of the cartel classifiers without knowing their identity?

To this end, let us first introduce some notation. Let the M classifiers be composed of a subset P of $(1-r)M$ "honest" classifiers and a subset C of rM malicious cartel classifiers. The honest classifiers satisfy the assumptions of the previous section: Each classifier attempts to correctly predict the truth with a balanced accuracy π_i, and different classifiers make independent errors. The cartel classifiers, in contrast, attempt to predict a different target labeling, $\mathbf{T}$. We assume that conditional on both the cartel's target and the true label, the classifiers in the cartel make independent errors. Namely, for all $i,j\in C$, and for any labels $a_i, a_j, Y, T \in \{-1,1\}$

$$\Pr\left[f_i(X)=a_i, f_j(X)=a_j \middle| T,Y\right] = \Pr[f_i(X)=a_i|T]\cdot \Pr\left[f_j(X)=a_j \middle| T\right]. \quad \textbf{[16]}$$

Similarly to the previous sections, we assume that the prediction errors of cartel and honest classifiers are also (conditionally) independent.

The following lemma, proven in *SI Appendix*, expresses the entries of the population covariance matrix Q in terms of the following quantities: the balanced accuracies of the M classifiers, the balanced accuracy π_c of the cartel's target with respect to the truth, and the balanced accuracies ξ_j of the rM cartel members relative to their target.

Lemma 2. *Given $(1-r)M$ honest classifiers and rM classifiers of a cartel C, the entries q_{ij} of Q satisfy*

$$q_{ij} = \begin{cases} 1-\mu_i^2 & i=j \\ (2\pi_i-1)(2\pi_j-1)(1-b^2) & i\in P, j\in P \\ (2\pi_i-1)(2\pi_c-1)(2\xi_j-1)(1-b^2) & i\in P, j\in C \\ (2\xi_i-1)(2\xi_j-1)(1-b^2) & i\in C, j\in C \end{cases}, \quad \textbf{[17]}$$

where $b\in(-1,1)$ is the class imbalance, as in Eq. **6**.

Next, the following theorem shows that in the presence of a single cartel the off-diagonal entries of Q correspond to a rank-two matrix. We conjecture that in the presence of k independent cartels, the respective rank is $(k+1)$.

Theorem 1. *Given $(1-r)M$ honest classifiers and rM classifiers belonging to a cartel, $0<r<1$, the off-diagonal entries of Q correspond to a rank-two matrix with eigenvalues*

$$\begin{aligned} \lambda_1 &= \lambda_P\cos^2\alpha + \lambda_C\sin^2\beta \\ \lambda_2 &= \lambda_P\sin^2\alpha + \lambda_C\cos^2\beta \end{aligned} \quad \textbf{[18]}$$

and eigenvectors

$$e_{1i} = \begin{cases} (2\pi_i-1)\cos\alpha & i\in P \\ (2\xi_i-1)\sin\beta & i\in C \end{cases} \qquad e_{2i} = \begin{cases} (2\pi_i-1)\sin\alpha & i\in P \\ (2\xi_i-1)\cos\beta & i\in C \end{cases}, \quad \textbf{[19]}$$

where

$$\lambda_P = (1-b^2)\sum_{j\in P}(2\pi_j-1)^2, \; \lambda_C = (1-b^2)\sum_{j\in C}(2\xi_j-1)^2 \quad \textbf{[20]}$$

and, with $k_1 = 2\pi_c - 1$, $k_2 = \lambda_C/\lambda_P$,

$$\alpha = \frac{1}{2}\arctan\left(\frac{k_1k_2}{k_2(1-2k_1^2)-1}\right), \qquad \beta = \frac{1}{2}\arctan\left(\frac{2k_1\sqrt{1-k_1^2}}{1-k_2-2k_1^2}\right).$$

As an illustrative example of *Theorem 1*, consider the case where the cartel's target is unrelated to the truth, i.e., $\pi_c = 1/2$. In this case $\alpha=\beta=0$, so $\lambda_1=\lambda_P$, $\lambda_2=\lambda_C$ and

$$e_{1i} = \begin{cases} 2\pi_i-1 & i\in P \\ 0 & i\in C \end{cases} \quad e_{2i} = \begin{cases} 0 & i\in P \\ 2\xi_i-1 & i\in C \end{cases}. \quad \textbf{[21]}$$

Next, according to Eq. **15** SML weighs each classifier by the corresponding entry in the leading eigenvector. Hence, if the cartel's target is orthogonal to the truth $(\pi_c = 1/2)$ and $\lambda_P > \lambda_C$, SML asymptotically ignores the cartel (*SI Appendix*, Fig. S3). In contrast, regardless of π_c, majority voting is affected by the cartel, proportionally to its fraction size r. Hence, SML is more robust than majority voting to the presence of such a cartel.

Application to Simulated and Real-World Datasets

The examples provided in this section showcase strengths and limitations of spectral approaches to the problem of ranking and combining multiple predictors without access to labeled data. First, using simulated data of an ensemble of independent classifiers and an ensemble of independent classifiers containing one cartel, we confirm the expected high performance of our ranking and SML algorithms. In the second part we consider the predictions of 33 machine learning algorithms as our ensemble of binary classifiers and test our spectral approaches on 17 real-world datasets collected from a broad range of application domains.

Simulations. We simulated an ensemble of $M = 100$ independent classifiers providing predictions for $S = 600$ instances, whose ground truth had class imbalance $b = 0$. To imitate a difficult setting, where some classifiers are worse than random, each generated classifier had different sensitivity and specificity chosen at random such that its balanced accuracy was uniformly distributed in the interval $[0.3, 0.8]$. We note that classifiers that are worse than random may occur in real studies, when the training data are too small in size or not sufficiently representative of the test data. Finally, we considered the effect of a malicious cartel consisting of 33% of the classifiers, having their own target labeling. More details about the simulations are provided in *SI Appendix*.

Ranking of classifiers. We constructed the sample covariance matrix, corrected its diagonal as described in *SI Appendix*, and computed its leading eigenvector $\hat{\mathbf{v}}$. In both cases (independent classifiers and cartel), with probability of at least 80%, the classifier with highest accuracy was also the one with the largest entry (in absolute value) in the eigenvector $\hat{\mathbf{v}}$, and with probability >99% its inferred rank was among the top five classifiers (*SI Appendix*, Fig. S4). Note that even if the test data of size $S = 600$

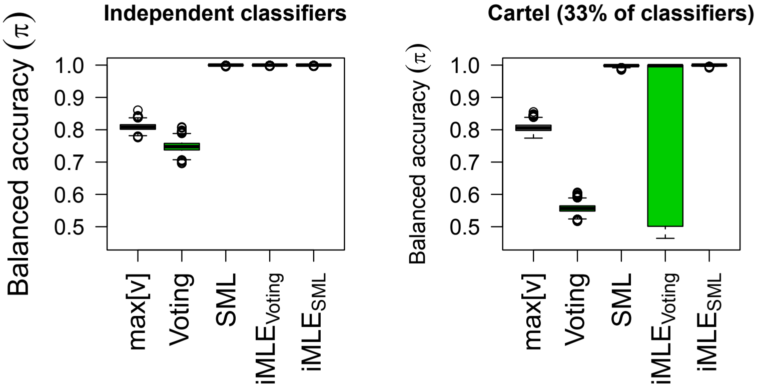

Fig. 1. Balanced accuracy of several classifiers: the classifier f_i with largest eigenvector entry, $i = \text{argmax}_j |v_j|$; majority voting; SML; and iMLE starting from SML or from majority voting. (*Left*) One hundred independent classifiers; (*Right*) 67 honest classifiers and 33 belonging to a cartel with target balanced accuracy of 0.5. The boxplots represent the distribution of balanced accuracies of 3,000 independent runs.

were fully labeled, identifying the best-performing classifier would still be prone to errors, because the estimated balanced accuracy has itself an error of $O(1/\sqrt{S})$.

Unsupervised ensemble learning. Next, for the same set of simulations we compared the balanced accuracy of majority voting and of SML. We also considered the predictions of these two metalearners as starting points for iterative EM calculation of the MLE (iMLE). As shown in Fig. 1, SML was significantly more accurate than majority voting. Furthermore, applying an EM procedure with SML as an initial guess provided relatively small improvements in the balanced accuracy. Majority voting, in contrast, was less robust. Moreover, in the presence of a cartel, computing the MLE with majority voting as its starting point exhibited a multimodal behavior, sometimes converging to a local maxima with a relatively low balanced accuracy.

A more detailed study of the sensitivity of SML and majority voting and their respective iMLE solutions versus the size of a malicious cartel with $\pi_c = 0.5$ is shown in *SI Appendix*, Fig. S5. As expected, the average balanced accuracy of all methods decreases as a function of the cartel's fraction r, and once the cartel's fraction is too large all approaches fail. In our simulations, both SML and iMLE initialized with SML were far more robust to the size of the cartel than either majority voting or iMLE initialized with majority voting. With a cartel size of 20%, SML was still able to construct a nearly perfect predictor, whereas the balanced accuracy of majority voting and iMLE initialized with majority voting were both far from 1. Interestingly, in our simulations, iMLE using SML as starting condition showed no significant improvement relative to the average balanced accuracy of SML itself.

Real Datasets. We applied our spectral approaches to 17 different datasets of moderate and large sizes from medical, biological, engineering, financial, and sociological applications. Our ensemble of predictions was composed of 33 machine-learning methods available in the software package Weka (25) (*Materials and Methods*). We split each dataset into a labeled part and an unlabeled part, the latter serving as the test data D used to evaluate our methods. To mirror our problem setting, each algorithm had access and was trained on different subsets of the labeled data (*SI Appendix*).

Fig. 2 and *SI Appendix*, Figs. S6–S8 show the results of different metaclassifiers on these datasets. Let us now interpret these results and explain the apparent differences in balanced accuracy between different approaches, in light of our theoretical analysis in the previous section.

In datasets where our assumptions are approximately satisfied, we expect SML, iMLE initialized with SML, and iMLE initialized with majority voting to exhibit similar performances. This is the case in the ACS data (Fig. 2, *Left*), and in all datasets in *SI Appendix*, Fig. S6. We verified that in these datasets Eq. **3** indeed holds approximately (*SI Appendix*, Table S3). In addition, in all these datasets, the corresponding sample covariance matrix of the 33 classifiers was almost rank-one with $\lambda_1(\hat{R})/\text{Trace}(\hat{R}) > 0.8$.

SI Appendix, Fig. S7 and Fig. 2, *Center* correspond to datasets where the median performance of the classifiers was only slightly above 0.5, with some classifiers having poor, even worse than random balanced accuracy. Interestingly, in these datasets, the covariance matrix between classifiers was far from being rank-one (similar to the case when cartels were present). The relative amount of variance captured by the first two leading eigenvalues $\lambda_1/\sum\lambda_j$ and $\lambda_2/\sum\lambda_j$ was, on average, 51.4% and 15.0%, respectively. In these datasets, SML seems to offer a clear advantage: Initializing iMLE with SML rather than with majority voting avoids the poor outcomes observed in the NYSE, AMEX, and PNS datasets.

Finally, the datasets in *SI Appendix*, Fig. S8 and in Fig. 2, *Right* are characterized by very sparse (ENRON) or high-dimensional (LASTFM) feature spaces $\mathcal{X}$. In these datasets, some instances were highly clustered in feature space, whereas others were isolated. Thus, in these datasets many classifiers made identical errors.

Remarkably, even in these cases, iMLE initialized with SML had an equal or higher median balanced accuracy than iMLE initialized with majority voting. This was consistent across all datasets, indicating that the SML prediction provided a better starting point for iMLE than majority voting.

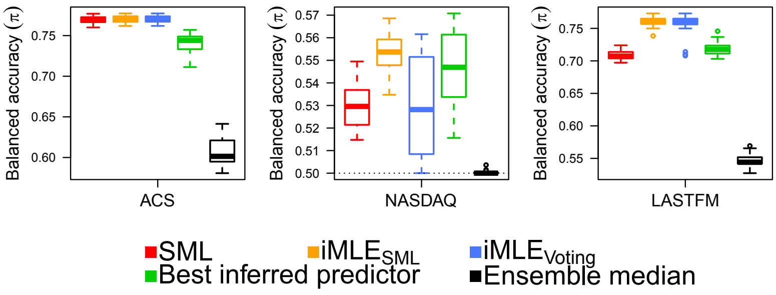

Fig. 2. Comparison between SML, iMLE from SML or from majority voting, the best inferred predictor, and the median balanced accuracy of all ensemble predictors on real-world datasets. The ACS dataset (*Left*) approximately satisfied our assumptions. In the NASDAQ dataset (*Center*) many predictors had poor performances. For the LASTFM data (*Right*), the predictors did not satisfy the conditional independence assumption. In all cases iMLE starting from SML had equal or higher balanced accuracy than iMLE starting from majority voting. The boxplots represent the distribution of balanced accuracies over 30 independent runs.

Summary and Discussion

In this paper we presented a spectral-based statistical analysis for the problems of unsupervised ranking and combining of multiple predictors. Our analysis revealed that under standard independence assumptions the off-diagonal of the classifiers covariance matrix corresponds to a rank-one matrix, whose eigenvector entries are proportional to the classifiers balanced accuracies. Our work gives a computationally efficient and asymptotically consistent solution to the classical problem posed by Dawid and Skene (15) in 1979, for which to the best of our knowledge only nonconvex iterative likelihood maximization solutions have been proposed (18, 26–29).

Our work not only provides a principled spectral approach for unsupervised ensemble learning (such as our SML), but also raises several interesting questions for future research. First, our proposed spectral-based SML has inherent limitations: It may be suboptimal for finite samples, in particular when one classifier is significantly better than all others. Furthermore, most of our analysis was asymptotic in the limit of an infinitely large unlabeled test set, and assuming perfect conditional independence between classifier errors. A theoretical study of the effects of a finite test set and of approximate independence between classifiers on the accuracy of the leading eigenvector is of interest. This is particularly relevant in the crowdsourcing setting, where only few entries in the prediction matrix $f_i(x_k)$ are observed. Although an estimated covariance matrix can be computed using the joint observations for each pair of classifiers, other approaches that directly fit a low-rank matrix may be more suitable.

Second, a natural extension of the present work is to multiclass or regression problems where the response is categorical or continuous, instead of binary. We expect that in these settings the covariance matrix of independent classifiers or regressors is still approximately low-rank. Methods similar to ours may improve the quality of existing algorithms.

Third, the quality of predictions may also be improved by taking into consideration instance difficulty, discussed, for example, in refs. 18 and 23. These studies assume that some instances are harder to classify correctly, independent of the classifier used, and propose different models for this instance difficulty. In our context, both very easy examples (on which all classifiers agree) and very difficult ones (on which classifier predictions are as a good as random) are not useful for ranking the different classifiers. Hence, modifying our approach to incorporate instance difficulty is a topic for future research.

Finally, our work also provides insights on the effects of a malicious cartel. The study of spectral approaches to identify cartels and their target, as well as to ignore their contributions, is of interest owing to its many potential applications, such as electoral committees and decision making in trading.

Materials and Methods

Datasets and Classifiers. We used 17 datasets for binary classification problems from science, engineering, data mining, and finance (*SI Appendix*, Table S1). The classifiers used are described in ref. 30 or are implemented in the Weka suite (25) (*SI Appendix*, Table S2).

Statistical Analysis and Visualization. Statistical analysis and visualization were performed using MATLAB (2012a; The MathWorks) and R (www.R-project.org). Additional information is provided in *SI Appendix*.

ACKNOWLEDGMENTS. We thank Amit Singer, Alex Kovner, Ronald Coifman, Ronen Basri, and Joseph Chang for their invaluable feedback. The Wisconsin breast cancer dataset was collected at the University of Wisconsin Hospitals by Dr. W. H. Wolberg and colleagues. F.S. is supported by the American–Italian Cancer Foundation. B.N. is supported by grants from the Israeli Science Foundation and from Citi Foundation. Y.K. is supported by the Peter T. Rowley Breast Cancer Research Projects (New York State Department of Health) and National Institutes of Health Grant R0-1 CA158167.

1. Tsai C-F, Hsiao Y-C (2010) Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decis Support Syst* 50:258–269.
2. Zweig J (January 8, 2011) Making sense of market forecasts. *Wall Street Journal* Available at http://online.wsj.com/news/articles/SB10001424052748703675904576064320900295678. Accessed January 2, 2014.
3. European Central Bank Survey of professional forecasters. Available at www.ecb.int/stats/prices/indic/forecast/html/index.en.html. Accessed December 29, 2013.
4. Federal Reserve Bank of Philadelphia Survey of professional forecasters. Available at www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/. Accessed December 29, 2013.
5. Monetary Authority of Singapore Survey of professional forecasters. Available at www.mas.gov.sg/news-and-publications/surveys/mas-survey-of-professional-forecasters.aspx. Accessed December 29, 2013.
6. Genre V, Kenny G, Meyler A, Timmermann AG (2010) Combining the forecasts in the ECB survey of professional forecasters: Can anything beat the simple average? (Social Science Research Network, Rochester, NY), SSRN Scholarly Paper ID 1719622.
7. Micsinai M, et al. (2012) Picking ChIP-seq peak detectors for analyzing chromatin modification experiments. *Nucleic Acids Res* 40(9):e70.
8. Wright FC, De Vito C, Langer B, Hunter A; Expert Panel on Multidisciplinary Cancer Conference Standards (2007) Multidisciplinary cancer conferences: A systematic review and development of practice standards. *Eur J Cancer* 43(6):1002–1010.
9. Cole, S, et al. (1978) Peer review in the national science foundation: Phase one of a study (Office of Publications, National Academy of Sciences, Washington, DC).
10. Margolin AA, et al. (2013) Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci Transl Med* 5:181.
11. Linstone H, Turoff Me (1975) *Delphi Method: Techniques and Applications* (Addison–Wesley, Reading, MA).
12. Stock JH, Watson MW (2006) Forecasting with many predictors. *Handbook of Economic Forecasting*, eds Elliott G, Granger C, Timmermann A (Elsevier, New York), Vol. 1, pp 515–554.
13. Timmermann A (2006) Forecast combinations. *Handbook of Economic Forecasting*, eds Elliott G, Granger C, Timmermann A (Elsevier, New York), Vol. 1, pp 135–196.
14. Degroot MH (1974) Reaching a concensus. *J Am Stat Assoc* 69:118–121.
15. Dawid AP, Skene AM (1979) *J R Stat Soc Ser C Appl Stat* 28:20–28.
16. Karger DR, Oh S, Shah D (2011) Budget-optimal crowdsourcing using low-rank matrix approximations. *Proceedings of the IEEE Allerton Conference on Communication, Control, and Computing* (IEEE, New York), pp 284–291.
17. Karger D R, Oh S, Shah D (2011) Iterative learning for reliable crowdsourcing systems. arXiv:1110.3564v4.
18. Whitehill J, et al. (2009) Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Adv Neural Inf Process Syst* 22:2035–2043.
19. Smyth P, Fayyad U, Burl M, Perona P, Baldi P (1994) Inferring ground truth from subjective labelling of venus images. *Advances in Neural Information Processing Systems*, eds Tesauro G, Touretzky DS, Leen TK (Neural Information Processing Systems Foundation, Denver), pp 1085–1092.
20. Sheng VS, Provost F, Ipeirotis PG (2008) Get another label? Improving data quality and data mining using multiple, noisy labelers. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Assoc Computing Machinery, New York), pp 614–622.
21. Welinder P, et al. (2010) The multidimensional wisdom of crowds. *Advances in Neural Information Processing Systems 23*, eds Lafferty J, Williams C, Shawe-Taylor J, Zemel R, Culotta A, (Neural Information Processing Systems, La Jolla, CA), pp 2424–2432.
22. Yan Y, et al. (2010) Modeling annotator expertise: Learning when everybody knows a bit of something. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pp 932–939.
23. Raykar VC, et al. (2010) Learning from crowds. *J Mach Learn Res* 11:12971322.
24. Dietterich TG (2000) Ensemble methods in machine learning. *Lecture Notes in Computer Science* (Springer, Berlin), Vol 1857, pp 1–15.
25. Witten IH, Frank E, Hall MA (2011) Data Mining: *Practical Machine Learning Tools and Techniques* (Morgan Kaufmann, Burlington, MA).
26. Jin R, Ghahramani Z (2003) Learning with multiple labels. *Advances in Neural Information Processing Systems*, eds Becker S, Thrun S, Obermayer K (MIT, Cambridge, MA), Vol 15, pp 897–904.
27. Lauritzen SL (1995) The EM algorithm for graphical association models with missing data. *Comput Stat Data Anal* 19(2):191–201.
28. Snow R, O'Connor B, Jurafsky D, Ng AY (2008) Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Assoc Computational Linguistics, Stroudsburg, PA), pp 254–263.
29. Walter SD, Irwig LM (1988) Estimation of test error rates, disease prevalence and relative risk from misclassified data: A review. *J Clin Epidemiol* 41(9):923–937.
30. Strino F, Parisi F, Kluger Y (2011) VDA, a method of choosing a better algorithm with fewer validations. *PLoS ONE* 6(10):e26074.