Dynamic Sensing: Better Classification under Acquisition Constraints

Oran Richman RORAN@TX.TECHNION.AC.IL
Department of Electrical Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel

Shie Mannor SHIE@EE.TECHNION.AC.IL
Department of Electrical Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel

Abstract

In many machine learning applications the quality of the data is limited by resource constraints (may it be power, bandwidth, memory, ...). In such cases, the constraints are on the average resources allocated, therefore there is some control over each sample's quality. In most cases this option remains unused and the data's quality is uniform over the samples. In this paper we propose to actively allocate resources to each sample such that resources are used optimally overall. We propose a method to compute the optimal resource allocation. We further derive generalization bounds for the case where the problem's model is unknown. We demonstrate the potential benefit of this approach on both simulated and real-life problems.

1. Introduction

Most machine learning methods take feature vectors as input. These features are often acquired using some noisy process resulting in less than optimal data quality. In many scenarios, the data quality depends on the resources allocated for the data acquisition process. Frequently, resources (power, memory, bandwidth,...) can be dynamically allocated while maintaining some global constraint on their average. Examples of such scenarios are:

- Due to bandwidth constraints the use of vector quantization (VQ) is popular (Linde et al., 1980). Such quantization can be viewed as adding noise to the input. One can dynamically switch VQ schemes while maintaining the average bandwidth rate.
- Due to power constraints, mobile devices often use lower than possible sampling rate. A feature's accuracy is often related to the sampling rate (Anderson, 2011), and therefore low sampling rate results in low accuracy. One can employ non-uniform sampling rate.
- Due to memory constraints, it is common practice to use sliding windows in spectral features calculation which causes spectral features to be inaccurate (Anderson, 2011). One may dynamically choose the length of the window to use.
- Due to computation constraints, features that require averaging are calculated using only part of the data (for example acquiring word frequencies from only part of the text). This causes these features to be inaccurately estimated. One can dynamically choose which part of the data to use.

Resources are usually allocated passively, such that all samples are acquired in the same way. We propose to actively allocate the resources across samples while maintaining the global resource constraints. In this way "easy" decisions require less resources. Therefore allowing to invest more resources in the "harder" cases. Figure 1 illustrates this approach in the case of support vector machine (SVM) classification. The figure shows the optimal resource allocation for the case of Gaussian noise (high amount of resources results in low noise). Far away from the decision boundary, few resources are needed, since even with large uncertainty the correct result is clear. Surprisingly, very near the decision boundary few resources are needed also. This is since the error will be close to 0.5 even when a lot of resources are allocated. Therefore, most of the resources should be allocated to samples which fall between those two extremes.

In this paper we propose a method for allocating resources in the *decision making* phase. We assume that special effort is made such that the training data is of the highest quality. During the test phase, however, resources are limited and should be allocated sparingly. This is often the case in applications where the number of samples to be classified is

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

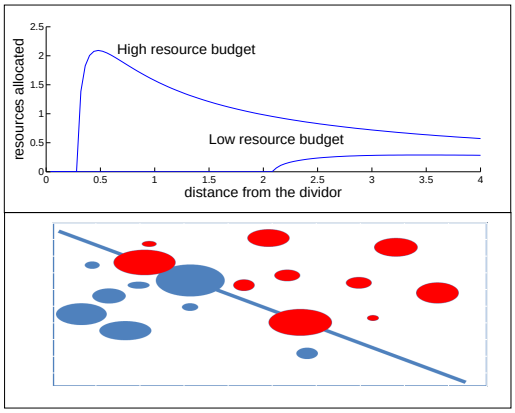

Figure 1. Illustration of optimal allocation for the case of SVM classification. Bigger circles indicates less resources and therefore more uncertainty

larger by several orders of magnitude from the training set size.

The contributions of this paper are threefold. First, we present a general model for problems of resource allocation in classification problems. To the best of our knowledge, this formulation is novel and such a problem has not been investigated before. Second, we define the resource allocation optimization problem and propose an efficient method for solving it. Lastly, we derive a bound on the error that results from not knowing the data distribution.

Related works Our approach shares the motivation with active classification (Heckerman et al., 1994). In active classification, a decision tree is used to acquire features on need. This tree is constructed such that the total cost is balanced with classification accuracy. Joint learning of the classifier and associated decision tree was considered (Greiner et al., 1996; Ji and Carin, 2007; Xu et al., 2014). In addition, similar schemes where features are also actively acquired during the learning phase (also known as budgeted learning) were investigated (Melville et al., 2004). Recent work had also explored a sequential approach where at each step a decision about which feature to acquire next has to be made. Both greedy (Gao and Koller, 2011; Saar-Tsechansky et al., 2009) and Dynamic programming algorithms (Kanani and McCallum, 2012) were considered. There is also a growing research interest in recent years in classifiers cascades (Vasconcelos and Saberian, 2010) where decisions are made sequentially and each stage employs more resources than its predecessor.

Our work differs from the above in several aspects. First, our model separates the system from the disturbance. This allows to better introduce prior knowledge about the disturbance structure. Second, we consider the decision space to be continuous and not discrete. This allows us to use new techniques. Our approach does not include heuristics and has little computational requirements in the decision making phase. Third, we propose a general probabilistic framework with a theoretical analysis of the overall classification scheme.

Another related field is that of active learning (Settles, 2010). In active learning, features are acquired for "free", however the learner can choose which labels to acquire. Choosing which label to acquire may result in a substantial improvement in the learning performance (Freund et al., 1997). Situations in which the label's quality can be controlled have also been investigated (Sheng et al., 2008). A related problem is that of active class selection (Lomasky et al., 2007) in which labels are known, however acquiring the data has a cost. As opposed to active learning, we are concerned with the quality of the features and not with the quality of the labels.

Other related work includes several methods that have been proposed in order to incorporate the knowledge that data are noisy in learning schemes; for examples (Xu et al., 2009), , (Trafalis and Gilbert, 2007), (El Ghaoui and Lebret, 1997). While these methods provide a way to deal with existing uncertainty, they do not try to actively manage it.

The paper is structured as follows: Section 2 formally defines the problem at hand. The main result of this paper is given in Section 3, where a general method to derive an optimal resource allocation is presented alongside an example. Section 4 explores the situation where the data distribution is unknown and provides a performance bound on the error resulting from the need to learn the distribution. Section 5 gives a taste of the method potential using simulation results on both a toy data set and real-life data. Section 6 concludes this paper with some final thoughts.

2. Model Formulation

We assume that samples $(x, y) \in (\chi \subset \mathbb{R}^d, \{-1, 1\})$ are generated i.i.d. from some joint distribution with a marginal density function $p(x)$. We assume that χ is closed and bounded. The data quality management (DQM) process is illustrated in Figure 2. A sample undergoes some coarse feature acquisition which produces a low quality feature vector $\tilde{x}$. We denote the resulting marginal density function $\tilde{p}(\tilde{x})$. We assume $\tilde{x}$ is also in χ. Both the underlying data model and the coarse acquisition model are known, namely $p(x)$, $\tilde{p}(\tilde{x})$ and $\tilde{p}(\tilde{x}|x)$ are assumed known. This feature vector is then re-acquired using resources allocated

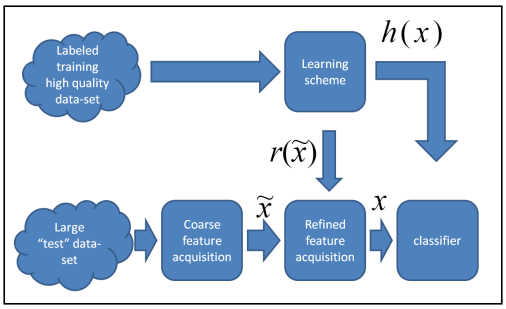

Figure 2. Data flow in Data quality management

using a pre-learned quality function $r(\tilde{x})$ to produce the feature vector x. This feature vector is then classified using a binary classifier $h(x)$. Both $h(x)$ and $r(\tilde{x})$ are previously learned using high quality training data. While learning $h(x)$ can be done using any learning algorithm, the main result of this paper is a method to derive $r(\tilde{x})$. We assume that the model that connects the allocated resources $r(\tilde{x})$ with the resulting disturbance in x is known. This assumption is quite reasonable. Examples include the influence of sampling rate on temporal features, sampling time of spectral features, power in communication and radar and many more (Anderson, 2011).

Denote by $P^+(x) = \mathbb{P}(Y = h(X)|X = x)$ and $P^-(x) = \mathbb{P}(Y \neq h(X)|X = x)$ the posterior performance measures of h. In addition, $\Delta P(x) = P^+(x) - P^-(x)$. The error which results from the disturbance δ generated using r resources can be stated as:

$$G(x,r) \triangleq \Delta P(x)\mathbb{P}(h(x+\delta(r)) \neq h(x)),$$
$$\overline{G}(\tilde{x},r) \triangleq \mathbb{E}_x(G(x,r)|\tilde{x}).$$

The partial derivatives in r is denoted as

$$g_x(r) = g(x,r) = \frac{-\partial G(x,r)}{\partial r},$$
$$\overline{g}(\tilde{x},r) = E_x(g(x,r)|\tilde{x}).$$

We assume that noise decreases performance; formally this assumption states that $\overline{G}(\tilde{x},r)$ is positive. We further assume that $g(x,r)$ is a positive continuous function and $g_x(r)$ is strictly decreasing for every $x \in \chi$ (and therefore, $\overline{g}_x(r)$ is also positive continuous and strictly decreasing). This mild assumption holds for many common distributions of disturbance.

We wish to find optimal allocation for the available resources constraints in the sense that classification results will be most similar to those obtained from optimal quality data. Namely solve the problem

$$\begin{array}{lll} \min_{r(\tilde{x})} L(r(\tilde{x})) & = & \mathbb{E}_{\tilde{x}}\overline{G}(\tilde{x},r(\tilde{x})) \\ s.t. \quad \mathbb{E}_{\tilde{x}}(r(\tilde{x})) & \leq & \beta \end{array}. \qquad (1)$$

Where β denotes the resource "budget" allocated. The maximal possible change rate is denoted by $g_{max} = \max_{x\in\chi} g(x,0)$.

An example Consider the special case of a linear SVM (Hsu et al., 2003), such that $h(x) = sign(w^\top x + b)$. As before, the data are corrupted by some noise. Assume that the noise is Gaussian with mean 0 and variance r^{-1}, where r denotes the resources allocated to the acquisition process. For simplicity, we assume $h(x)$ is a proper classifier. Now the error resulting from the disturbance is,

$$G(x,r) = \mathbb{P}(h(x+\delta(r)) \neq h(x)) = \Phi(|w^\top x + b|\sqrt{r}),$$

where Φ denotes the cumulative distribution function of the standard normal distribution. Further,

$$g_x(r) = \frac{|w^\top x + b|}{\sqrt{2\pi r}} e^{\frac{-(w^\top x+b)^2 r}{2}}.$$

3. Finding the Optimal Resource Allocation

We now move on to state the main result of this paper. The next theorem shows that the optimal resource allocation can be derived by solving an equations set that is equivalent to problem (1).

Theorem 1. If for every $\tilde{x} \in \chi$, $\tilde{p}(\tilde{x}) > 0$. Then for some $\lambda > 0$ a unique solution for problem (1) is given by

$$\begin{array}{lll} r(\tilde{x}) & = & \begin{cases} 0 & \overline{g}(\tilde{x},0) \leq \lambda \\ \overline{g}_{\tilde{x}}^{-1}(\lambda) & \overline{g}(\tilde{x},0) > \lambda \end{cases} \\ \beta & = & \int_\chi \tilde{p}(x) r(x) dx \end{array}. \qquad (2)$$

Where g^{-1} denote the inverse function of g. Note that solving this equations set will provide the desired λ.

Proof. The proof consists of four parts. First, we will show that a solution exists. Second, we will show that (2) is well defined. Third, we will show that (2) meets the necessary conditions for an optimum (the Euler equation; for more information see (Gelfand et al., 2000). Finally, we will show that no other solution can solve the Euler equation. Since the problem is written now only in terms of $\tilde{x}$, for ease of reading, we will use throughout the proof x instead of $\tilde{x}$ and in $g(x,r)$ instead of $\overline{g}(\tilde{x},r)$.

Part 1 Since $L(r(x))$ is bounded and continuous, in order to show that a solution exists it suffices to show that the set of possible solutions $\{r(x)\}$ can be bounded. Denote by $\hat{r}(x)$ the optimum of (1) and assume by contradiction that it is not bounded. Namely, that for every chosen $\alpha > 2\beta$

$$\Delta x = \mathbb{P}(\hat{r}(x) > \alpha) > 0.$$

On the other hand, from the second part of equation (2) it can be seen that $\Delta x < \frac{\beta}{\alpha}$.

Define a new quality function

$$\tilde{r}(x) = \begin{cases} 0 & \hat{r}(x) > \alpha \\ \hat{r}(x) + \Delta x\alpha & \hat{r}(x) \leq \alpha \end{cases}. \quad (3)$$

It is clear that (3) meets the resource constraints. Now,

$$\Delta L \triangleq L(\hat{r}) - L(\tilde{r}) = \mathbb{E}[G(x,\hat{r}(x)) - G(x,0); \hat{r}(x) > \alpha] + \mathbb{E}[G(x,\hat{r}(x)) - G(x,\hat{r}(x) + \Delta x\alpha); \hat{r}(x) \leq \alpha] > -\Delta x + \mathbb{E}[G(x,\hat{r}(x)) - G(x,\hat{r}(x) + \Delta x\alpha); \hat{r}(x) \leq 2\beta] > -\Delta x + \mathbb{E}[g(x,\hat{r}(x) + \Delta x\alpha)\Delta x\alpha; \hat{r}(x) \leq 2\beta] > -\Delta x + \tfrac{1}{2} g_{min}(3\beta)\Delta x\alpha,$$

where $g_{min}(r) = \min_{x\in\chi} g(x,r)$. The last inequality holds since $g(x,r)$ is decreasing in r and since $\mathbb{E}(\hat{r}(x)) = \beta$ implies that $\mathbb{P}(\hat{r}(x) < 2\beta) > \frac{1}{2}$. Clearly, for α big enough $\Delta L > 0$ which contradict the assumption that $\hat{r}$ is the optimum.

Part 2 We will now show that (2) is well defined, meaning that for every $\beta > 0$ a unique solution for (2) exists. We start by noting that since $g_x(r)$ is strictly decreasing with bounded integral then $g_x^{-1}(\lambda)$ is defined for $\lambda \in (0\ g_x(0)]$. Moreover, $g_x^{-1}(\lambda)$ is strictly decreasing in λ and $\lim_{\lambda\to 0} g_x^{-1}(\lambda) = \infty$.

We define the following sets:

$$I^+(\lambda) = \{x| \quad g(x,0) \geq \lambda\},$$
$$I^-(\lambda) = \{x| \quad g(x,0) < \lambda\}.$$

We further define

$$\beta(\lambda) = \int_{\chi} p(x)r(x)dx = \int_{I^+(\lambda)} p(x)g_x^{-1}(\lambda)dx.$$

The following limits are easy to show:

$$\lim_{\lambda\to g_{max}} \beta(\lambda) = \lim_{\lambda\to g_{max}} \int_{I^+(\lambda)} p(x)g_x^{-1}(\lambda)dx = 0,$$
$$\lim_{\lambda\to 0} \beta(\lambda) = \lim_{\lambda\to 0} \int_{I^+(\lambda)} p(x)g_x^{-1}(\lambda)dx = \infty.$$

If $\beta(\lambda)$ is strictly decreasing in λ than, for every β there will be a unique λ which meet the constraint $\beta(\lambda) = \beta$. In order to show that $\beta(\lambda)$ is strictly decreasing we assume, without loss of generality, that $\tilde{\lambda} < \lambda$. From the definition of $I^+(\lambda)$ it is obvious that $I^+(\lambda) \subseteq I^+(\tilde{\lambda})$. Moreover,

$$r(x) = \begin{cases} 0 \leq \tilde{r}(x) & g(x,0) \leq \lambda \\ g_x^{-1}(\lambda) < g_x^{-1}(\tilde{\lambda}) = \tilde{r}(x) & g(x,0) > \lambda > \tilde{\lambda} \end{cases}$$

Now,

$$\beta(\lambda) = \int_{\chi} p(x)r(x)dx = \int_{I^+(\lambda)} p(x)r(x)dx < \int_{I^+(\lambda)} p(x)\tilde{r}(x)dx \leq \int_{I^+(\tilde{\lambda})} p(x)\tilde{r}(x)dx = \beta(\tilde{\lambda}),$$

and $\beta(\lambda)$ is strictly decreasing in λ .

Part 3 For the next part of the proof we note that from calculus of variations (Gelfand et al., 2000) it is known that the optimum must satisfy the Euler equation. The Euler equation for problem (1) is given by

$$\begin{array}{llll} \text{if} & r(x) = 0 & \text{then} & g(x,0) \leq \lambda \\ \text{if} & r(x) > 0 & \text{then} & g(x,r(x)) = \lambda \\ \beta & & = & \int_{\chi} p(x)r(x)dx. \end{array} \quad (4)$$

It is easy to verify that a solution to (2) solves equations (4).

Part 4 It is now left to show that no other solution can solve (4). Since we have already shown that $\beta(\lambda)$ is strictly decreasing showing that every solution of the Euler equation is of the form (2) will prove the theorem. In order to show that, we will note the following set of conditions: If $g(x,0) > \lambda$ then from the first condition in (4) $r(x) > 0$. Using the second condition yields $r(x) = g^{-1}(\lambda)$. If $g(x,0) < \lambda$ then $g(x,r(x)) \leq g(x,0) < \lambda$. Using the second condition in (4) yields $r(x) = 0$ Finally, if $g(x,0) = \lambda$ then $r(x) = 0$. This is since that if $r(x) > 0$ then $g(x,r(x)) < g(x,0) = \lambda$ which is a contradiction to the second condition in (4). □

Remark 1. The assumption that $\tilde{p}(x) > 0$ is technical and can be relaxed. When $\tilde{p}(x) = 0$ the resources allocated have no meaning and can receive any value. This may cause $r(x)$ to be non-continuous in x, that complicates the problem technically. It can be defined that whenever $\tilde{p}(x) = 0$, $r(x) = 0$ to avoid technicalities.

Remark 2. The assumption that χ is bounded is also technical and can be replaced with milder conditions. It is used only to ensure that $\mathbb{E}(r(x))$ exists and finite where r is the optimal allocation given by (2).

Revisiting the example shown earlier (linear classifier with Gaussian noise) we demonstrate how Theorem 1 can be used to derive the optimal resource allocation for a specific model. Experimental results matching this model can be found in Section 5.

As a reminder, $g_x(r) = \frac{|w^\top x+b|}{\sqrt{2\pi r}} e^{\frac{-(w^\top x+b)^2 r}{2}}$.

We assume that the coarse measurement $\tilde{x}$ is corrupted by śome Gaussian noise with variance R^2. Using (2) it is not

difficult to show that the optimal resource allocation can be obtained by solving the equations set:

$$\begin{aligned}\lambda &= \int\limits_{y\in\chi} \frac{1}{\sqrt{2\pi}R}\exp\frac{|y-\tilde{x}|^2}{2R^2}\frac{|w^\top y+b|}{\sqrt{r(\tilde{x})}}e^{\frac{-(w^\top y+b)^2 r(\tilde{x})}{2}}dy \\ \beta &= \mathbb{E}_{\tilde{x}}[r(\tilde{x})]\end{aligned}.$$

Note that knowing the correct λ allows the calculation of $r(\tilde{x})$, therefore a solution to the equations set can be found using single variable search methods. Note also that contrary to the assumptions made earlier $\tilde{x}$ is not bounded. This however, is of little practical implication as explained by Remark 2.

4. Unknown Data Distribution

Let us now turn our attention to the more practical case where the classifier h is known but, the data distribution is unknown. The model should be learned from a finite training sample of coarsely acquired data $\{\tilde{X}_i, Y_i\}_{i=1,\ldots,N}$. Note that although the distributions are unknown, the loss functions $G(x,r)$ and $\overline{G}(\tilde{x},r)$ are known. We propose to approximate the uncertainty functional $r(x)$ by using K Radial basis functions (RBF) such that

$$r(x) = \sum_{i=1}^{K}\alpha_i\phi_i(x),$$

where $\{\phi_i\}_{i=1,\ldots,K} = \phi(|x-x_i|)$ for some set $\{x_i\}_{i=1,\ldots,K}$. We assume that $\phi_i(x)$ is bounded from below by some constant $\phi_{min} > 0$ almost everywhere. Note that this assumption always holds if x is bounded. Since $r(x)$ is always positive we can limit ourselves to choosing the vector α such that $\forall i\ \alpha_i \geq 0$. Now (1) can be substituted by:

$$\begin{aligned}&\min_{(\alpha_1,\ldots,\alpha_K)} R_{emp}(r(x)) = \tfrac{1}{N}\sum_{i=1}^{N}\bar{G}(\tilde{X}_i, \sum_{j=1}^{K}\alpha_j\phi_j(\tilde{X}_i)) \\ &s.t\ \ \beta(r) = \tfrac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\alpha_j\phi_j(\tilde{X}_i) \leq \beta, \\ &\qquad \alpha_i \geq 0 \quad \forall i \in \{1,\ldots,K\}.\end{aligned} \tag{5}$$

Since $\phi_j(x) > 0$ and G is convex, problem (5) is convex and can be solved using conventional methods. We will define the hypothesis space $H_{N,K}$ as the set of vectors $\{\alpha_1,\ldots,\alpha_K\}$ for which there exists a sample $S = (X_i,Y_i)_{i=1,\ldots,N}$ such that $\{\alpha_1,\ldots,\alpha_K\}$ solves (5).

We wish now to bound the generalization error that results from the need to learn the distributions. For that purpose we will use the Rademacher complexity that measures the complexity of a hypothesis space with respect to a certain training set. The generalization error can be bounded using the Rademacher complexity (we use the same notation as Surhone et al., 2010). We will therefore establish a bound on the Rademacher complexity. We also establish a bound on the actual average resource consumption. The actual resource consumption can differ from the constrains since it depends on the coarse data $\{\tilde{X}_i\}_{i=1,\ldots,N}$. In cases where the constraints are rigid proper slacks should be taken.

Denote by $\hat{r}_S(x)$ the solution of (5) for sample S and

$$l(X_i) = \bar{G}(X_i, \sum_{j=1}^{K}\alpha_j\phi_j(X_i)).$$

The bound can be stated as:

Theorem 2. For every natural number $N > 0$, and every sample $S = (X_i,Y_i)_{i=1,\ldots,N}$ the Radamacher complexity is bounded by

$$R(l\circ H_{N,K}, S) \leq \frac{\beta}{\phi_{min}}\bar{g}_{max}\sqrt{\frac{2\log K}{N}}. \tag{6}$$

Moreover $\forall\epsilon > 0$, and for any other sample S'

$$\mathbb{P}(|\beta^{S'}(\hat{r}_S(x)) - \beta|) > \epsilon) \leq e^{-2N(\frac{\epsilon\phi_{min}}{\beta})^2}. \tag{7}$$

Where

$$\beta^{S'}(\hat{r}_S(x)) = \frac{1}{N}\sum_{i=1}^{N}\hat{r}_S(X_i^{S'})$$

Proof. The proof is rather standard, the set of possible hypotheses is bounded and Radamacher calculus is used to derive the desired bound on the generalization error. From (5) it is clear that $||\alpha||_1 \leq \frac{\beta}{\phi_{min}}$.

We note that for every $\alpha,\gamma > 0$

$$|\bar{G}(x,\alpha) - \bar{G}(x,\gamma)| < g(x,0)|\alpha-\gamma| < \bar{g}_{max}|\alpha-\gamma|.$$

Therefore,

$$R(l\circ H_{N,K}, S) \leq \bar{g}_{max}R(l'\circ H_{N,K}, S),$$

where

$$l' = \sum_{j=1}^{K}\alpha_j\phi_j(X_i) =< \alpha, (\phi_1(X_i),\ldots,\phi_K(X_i)) > .$$

$< \circ,\circ >$ is used to denote the inner product.

The problem can now be stated as an L1 regression problem and it is known (Surhone et al., 2010) that

$$\begin{aligned}&R(l'\circ H_{N,K}, S) \leq \\ &\frac{\beta}{\phi_{min}}\max_i|(\phi_1(X_i),\ldots,\phi_K(X_i))|_\infty\sqrt{\frac{2\log K}{N}} \leq \\ &\qquad\qquad \frac{\beta}{\phi_{min}}\sqrt{\frac{2\log K}{N}},\end{aligned}$$

which conclude the proof of the first part of the theorem .

The bound on $\beta(\hat{r}(x))$ is derived using Hoeffding's inequality noticing that $\beta(\hat{r}(x))$ is bounded between 0 and $\frac{\beta}{\phi_{min}}$. □

Remark 3. In most cases $\bar{G}(\tilde{x}, r)$ is unknown, in contrary to the assumptions made. It can however be approximated without knowing the distribution of x . The approximation uses the fact that the relation between $\tilde{x}$ and x is local. Therefore, in most cases one can omit the prior and assume that $x = \tilde{x} - \delta$ where the distribution of δ is known. In a similar fashion $\Delta P(x)$ can be assumed to equal 1 over all the features space. This causes the algorithm to "waste" more resources on some samples than it should but in cases where the classifier is good this effect is small.

This approximation had been used in the simulation presented in Section 5 and provided good results.

5. Simulation Results

We tested our method on three data-sets. The first is a synthetic toy data-base. The second is the IRIS database from the UCI repository, in this database noise was added artificially. The third is a speaker verification corpus. This simulates a real-life scenario where noise is the result of reducing recording's length. In all tests our method (DQM) provided significant benefit over uniform allocation of resources

Toy database We created a toy-data set composed of 3000 linearly separable samples in $\mathbb{R}^4$, generated such that the distance from the separator is uniformly distributed between 0 to 3. The coarse samples $\tilde{x}$ were generated from the raw samples by adding zero mean Gaussian noise with standard deviation 0.33. Measurement noise was taken to be Gaussian with zero mean and variance $[0.33^2 + r(\tilde{x})]^{-1}$, where $r(\tilde{x})$ is the resources allocated to sample $\tilde{x}$. The problem is therefore similar to the problem presented earlier.

The method was then tested with a variety of resource budgets and compared to uniformly distributing the resources between samples. Figure 3 shows the performance measured, averaging over 100 consecutive runs. Note that in each run the noise generated for both resource allocation schemes is uncorrelated. The figure shows the classification error as a function of the noise variation. For DQM the noise variation refers to the variation associated with the average resource consumption. On each run we have calculated the ratio between the error for uniform allocation and the error for DQM (this ratio is known as lift). Figure 3 shows the lift for a total resource budget which matches an average variation of 0.08. It can be seen that the benefit of using DQM is significant. The average lift is 1.31 with

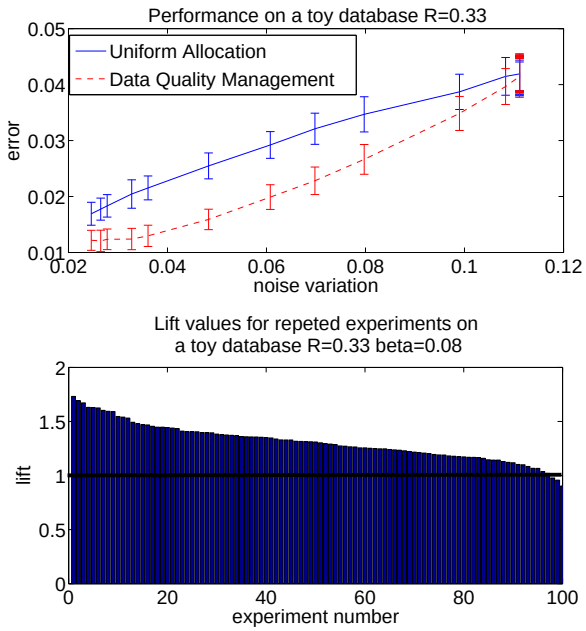

Figure 3. Performance on toy database

standard deviation of 0.17.

The improvement is however much lower for very high and very low resource budgets. In the former, since there is little room for improvement and in the latter, since little resources do little benefit no matter how they are divided.

IRIS dataset We tested our method also on the well known "IRIS" data-set taken from the UCI data repository (Bache and Lichman, 2013). The data-set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. There are 4 features which are the length and width of the sepal and the petal. We used this data-set to solve the binary classification problem of distinguishing Iris Virginica from all the rest. This problem is not linearly separable. We have chosen $r(x)$ to be the bandwidth needed to transfer the picture of the iris. The quantization noise generated from the picture resolution is proportional to the LSB (least significant bit) in each axis, which is inversely proportional to $\sqrt{r(x)}$. Therefore the setting match the setting that was presented on the toy data-set. Tests were conducted again with coarse acquisition standard deviation of 0.33. Since the data set is rather small, error was calculated by averaging over 20 different generated noise vectors. This creates the equivalent of 3000 items dataset. All experiments were repeated 100 times. Figure 4 shows the performance measured averaging over the 100 consecutive runs. Figure 4 shows lift values for a total resource budget that matches an average variation of 0.08. The benefit of using DQM is smaller than in the toy dataset. The average lift is 1.09 with standard deviation of 0.07. The re-

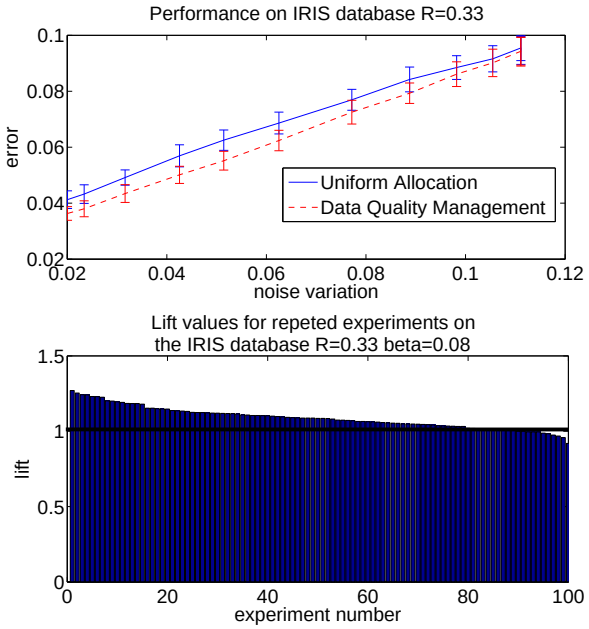

Figure 4. Performance on "IRIS" database

duced benefit is due to the structure of this particular data-set. Samples in this data-set reside in fairly tight clusters with similar distances from the divider. This causes the optimal solution to be close to uniformly allocation of resources.

Speaker identification In the two experiments shown so far artificial noise was added to the features. In order to test the method on a more realistic setting we further tested it on the task of speaker identification. Speaker identification classifiers often use short-time spectral features (Bimbot et al., 2004). Such feature's "noise level" is related to the length of the voice recordings used. Speaker identification had been studied extensively and is required for many applications (Bimbot et al., 2004). Due to the high computational cost required to extract the acoustic features, resources are often a concern. For example, imagine a mobile application that records an activity log of your day, whenever you meet someone it will open the microphone and try to recognize the other speakers in the conversation. This should be done using as little power as possible. Since the dominant part of computation resources used is for the recording and extraction of acoustic features, using shorter recordings means less power. Both resources used for recording and resources used for extraction of acoustic features are proportional to the length of the recording used.

We used the MIT Mobile Device Speaker Verification Corpus (Ram Woo and Hazen, 2006). This corpus contains short 2-words recordings collected using mobile devices in variable environment conditions. We have investigated the problem of distinguishing one speaker (for which we have 54 recordings) from all the other speakers (2160 "impostors" recording of 40 different speakers). For each sample, features extracted from the full recording ($\approx 3sec$) were regarded as "perfect". We refer to the difference between features generated using part of the recording to the "perfect" features as noise. Resources are controlled using a "sampling factor" which is inversely proportional to the portion of recording used. For example, a sampling factor of 2 means that only the first half of the recording is processed.

We implemented a simple text independent speaker identification engine based on SVM classification. SVM is known to provide good performance in this task (Fenglei and Bingxi, 2001). Mel-frequency scale cepstral coefficient (MFCC) are extracted from each recording[1]. This is done using a 25msec window with 10msec interval. Those coefficients are time-averaged, resulting in a vector of 13 features. The classifier is trained using the "perfect" features. The "noise" was modelled as Gaussian with mean 0 and standard deviation $\sigma \propto \sqrt{s-1}$ where s is the sampling factor. This modelling was done empirically and Figure 5 shows that this is quite reasonable. It can be seen that each feature reacts differently to the reduction in the amount of data. However, the distance from the dividing hyperplane roughly obeys this model. Notice that since the noise is not spherical the model used depend on the dividing hyperplane. Since the differences are not large, in practice one may use the same model for all dividing hyperplanes (persons to be identified).

From the database, 100 sets of samples were randomly chosen, each containing all 54 "positive" samples as well as 1000 "impostor" samples. For each set, an SVM classifier was trained using the maximal amount of data available. Then, coarse acquisition was preformed using 1/3 of the available data. The classifier performance was evaluated for different sampling factors. Resources were allocated both uniformly and by DQM. Averages and standard deviations was taken over the 100 datasets.

The results obtained can be seen in Figure 6. The figure also presents the lift values for a resource budget corresponding to processing half of the data. The average lift in this setting is 1.204 with standard deviation of 0.1 . It can be seen that the method provided significant benefit when there are enough resources to "make a difference". As can be seen on Figure 6, the method may provide $\approx 20\%$ resource reduction for the same desired performance level. Two observations are worth mentioning:

- When noise level is low, performance may improve by adding more noise. This is due to the fact that,

[1]We used HTK MFCC matlab package

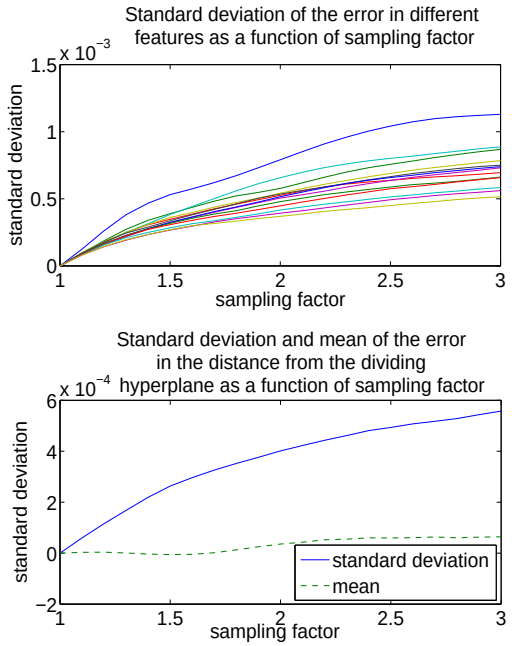

Figure 5. The effect of using a portion of the recording on the features values. Noise level as a function of sampling factor

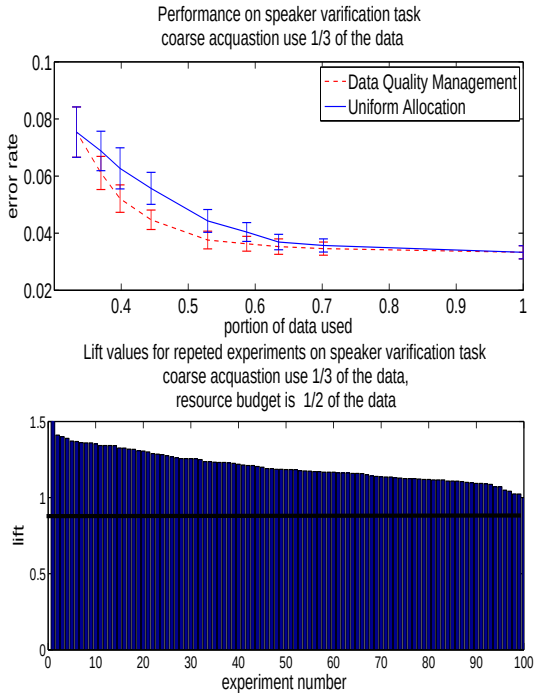

Figure 6. Performance in speaker verification task

contrary to the assumption, the classifier is not proper. With small amount of noise some samples "suffer" while other benefit, resulting in indifference to noise. When noise is significant this effect is negligible.

- While "noise" is assumed to be zero mean, as can be seen in Figure 5, in practice it is not. In low noise environments this bias is not negligible resulting in unexpected influence of the sampling factor on the classifier performance. This can possibly be compensated by incorporating some bias correction into the classifier. This bias may be the result of unvoiced segments in the recording, an issue that can be addressed by preprocessing.

6. Conclusion

In this work we presented a novel setting where the data's quality can be controlled by active allocation of resources to each sample. We believe that in many scenarios, careful allocation of resources can substantially improve performance. The improvement will be larger in cases where there is much diversity in classification "difficulty". However, scenarios in which the needed accuracy is more or less uniform will show little improvement. Such diversity is arguably common in real-life problems where systems are designed to meet performance in the worst case, while the average case requires much fewer resources.

There are two natural directions we would like to suggest for further research:

- We have chosen the loss function to be the expected error. One can use any other loss function. This can be beneficial, for example, in cases where the data-set is small and therefore the actual error may differ significantly from the expectation. As long as the loss function is continuous, strictly decreasing in r and convex in r, the results presented in this paper should hold.

- We have taken $\tilde{x}$ to be some coarse measurement of x. This can be substituted by any other measurement as long as the relationship $p(x|\tilde{x})$ is known. Note that $\tilde{x}$ and x do not even need to be within the same space. As an example, recall from the previous section the mobile activity log application with speaker recognition capability. In the model presented earlier, a short recording was used as a coarse measurement. Instead, $\tilde{x}$ may be the location of the user. Different locations will exhibit different class distributions and different background noise levels. Similar methods to the one presented can be used to derive $r(\tilde{x})$.

Acknowledgments

This work was partially supported by the Israel Science Foundation (ISF under contract 920/12)

References

T. W. Anderson. *The statistical analysis of time series*, volume 19. John Wiley & Sons, 2011.

K. Bache and M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds. A tutorial on text-independent speaker verification. *EURASIP journal on applied signal processing*, 2004:430–451, 2004.

L. El Ghaoui and H. Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18(4):1035–1064, 1997.

H. Fenglei and W. Bingxi. Text-independent speaker recognition using support vector machine. In *Info-tech and Info-net, 2001. Proceedings. ICII 2001-Beijing. 2001 International Conferences on*, volume 3, pages 402–407. IEEE, 2001.

Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine learning*, 28(2-3):133–168, 1997.

T. Gao and D. Koller. Active classification based on value of classifier. In *Advances in Neural Information Processing Systems*, pages 1062–1070, 2011.

I. Gelfand, S. Fomin, and R. Silverman. *Calculus of Variations*. Dover Books on Mathematics. Dover Publications, 2000. ISBN 9780486414485.

R. Greiner, A. J. Grove, and D. Roth. Learning active classifiers. In *ICML*, pages 207–215. Citeseer, 1996.

D. Heckerman, J. Breese, and K. Rommelse. Troubleshooting under uncertainty. *Communications of the ACM*, pages 121–130, 1994.

C.-W. Hsu, C.-C. Chang, C.-J. Lin, et al. Technical report, a practical guide to support vector classification, 2003.

S. Ji and L. Carin. Cost-sensitive feature acquisition and classification. *Pattern Recognition*, 40(5):1474–1485, 2007.

P. H. Kanani and A. K. McCallum. Selecting actions for resource-bounded information extraction using reinforcement learning. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 253–262. ACM, 2012.

Y. Linde, A. Buzo, and R. M. Gray. An algorithm for vector quantizer design. *Communications, IEEE Transactions on*, 28(1):84–95, 1980.

R. Lomasky, C. E. Brodley, M. Aernecke, D. Walt, and M. Friedl. Active class selection. In *Machine learning: ECML 2007*, pages 640–647. Springer, 2007.

P. Melville, M. Saar-Tsechansky, F. Provost, and R. Mooney. Active feature-value acquisition for classifier induction. In *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*, pages 483–486. IEEE, 2004.

A. P. Ram Woo and T. J. Hazen. The MIT mobile device speaker verification corpus: Data collection and preliminary experiments. In *Proceedings of Odyssey 2006, The Speaker and Language Recognition Workshop*, 2006.

M. Saar-Tsechansky, P. Melville, and F. Provost. Active feature-value acquisition. *Management Science*, 55(4): 664–684, 2009.

B. Settles. Active learning literature survey. *Technical report 1648, University of Wisconsin, Madison*, 2010.

V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622. ACM, 2008.

L. Surhone, M. Timpledon, and S. Marseken. *Rademacher Complexity*. VDM Publishing, 2010. ISBN 9786131121159.

T. Trafalis and R. Gilbert. Robust support vector machines for classification and computational issues. *Optimisation Methods and Software*, 22(1):187–198, 2007.

N. Vasconcelos and M. J. Saberian. Boosting classifier cascades. In *Advances in Neural Information Processing Systems*, pages 2047–2055, 2010.

H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *The Journal of Machine Learning Research*, 10:1485–1510, 2009.

Z. Xu, M. J. Kusner, K. Q. Weinberger, M. Chen, and O. Chapelle. Classifier cascades and trees for minimizing feature evaluation cost. *The Journal of Machine Learning Research*, 15(1):2113–2144, 2014.