Collaborative information acquisition for data-driven decisions

Danxia Kong · Maytal Saar-Tsechansky

Received: 17 November 2012 / Accepted: 13 November 2013 / Published online: 14 December 2013
© The Author(s) 2013

Abstract Data-driven predictive models are routinely used by government agencies and industry to improve the efficiency of their decision-making. In many cases, agencies acquire training data over time, incurring both direct and opportunity costs. Active learning can be used to acquire particularly informative training data that improve learning cost-effectively. However, when multiple models are used to inform decisions, prior work on active learning has significant limitations: either it improves the accuracy of predictive models without regard to how accuracy affects decision making or it addresses only decisions informed by a single predictive model. We propose that decisions informed by multiple models warrant a new kind of Collaborative Information Acquisition (CIA) policy that allows multiple learners to reason collaboratively about informative acquisitions. This paper focuses on tax audit decisions, which affect a vital revenue source for governments worldwide. Because audits are costly to conduct, active learning policies can help identify particularly informative audits to improve future decisions. However, existing active learning models are poorly suited to audit decisions, because audits are best informed by multiple predictive models. We develop a CIA policy to improve the decisions the models inform, and we demonstrate that CIA can substantially increase sales tax revenues. We also demonstrate that the CIA policy can improve decisions to target directly individuals in a donation campaign. Finally, we discuss and demonstrate the risks for decision making of the naive use of existing active learning policies.

Keywords Active learning · Predictive modeling · Decision making · Audit selection · Direct targeting · Tax audit

Editors: Cynthia Rudin and Kiri Wagstaff.

D. Kong
2310 W 9th Street, Unit 1, Austin, TX 78703, USA
e-mail: kongdanxia@utexas.edu

M. Saar-Tsechansky (✉)
Information, Risk and Operations Management, McCombs School of Business, The University of Texas
at Austin, 1 University Station, Austin, TX 78712, USA
e-mail: maytal@mail.utexas.edu

1 Introduction: active learning and decision making

Data-driven predictive models are routinely used to improve the decisions of government agencies and industries. However, the reliability of a data-driven model is limited by the information available from prior learning experiences, and additional experiences may be costly to acquire. Active learning can improve predictive modeling cost-effectively by identifying and acquiring particularly informative learning experiences. However, existing active learning approaches have significant limitations when multiple models are used to inform decisions. Much prior work on active learning does not consider the decisions the models inform. Consequently, this work aims to improve the accuracy of predictive models without regard to how accuracy affects the expected utility of the decision maker (e.g., Cohn et al. 1994; Abe and Mamitsuka 1998; Krogh and Vedelsby 1995; Lewis and Gale 1994; Reichart et al. 2008; Roy and McCallum 2001). Other work addresses only decisions informed by a single predictive model (Saar-Tsechansky and Provost 2007).

In this paper, we address decisions informed by *multiple* predictive models of different kinds, and we propose a new type of *Collaborative Information Acquisition* (CIA) policy. CIA allows *multiple* learners to reason collaboratively about opportunities to acquire informative learning experiences that cost-effectively improve the *decisions* the models inform. To our knowledge, no existing information acquisition policy accomplishes this objective.

We develop CIA in the context of the selection of sales tax audits. The sales tax is a primary revenue tool for most U.S. states, at times accounting for 50 % of a state budget (Cornelius 2012; EOI 2012). However, tax avoidance is a significant problem: a recent IRS estimate suggests that underreporting of income accounts for \$376 billion in lost revenues at the federal level alone (IRS 2012). This gap can be decreased by efficient audit selection that maximizes sales tax revenue. The decision-theoretic optimal choice, which corresponds to the action with the highest expected utility, is informed by two models, the estimated probability of non-compliance and the estimated revenue that can be recovered in the case of non-compliance. Both models are necessary to evaluate the audit revenue in expectation and they can each benefit from the costly acquisition of informative audits. However, a different audit might be more beneficial to each of the learners at any given time. CIA allows the learners to prioritize information acquisition collaboratively so as to improve the decisions they inform.

Using real sales tax data from the state of Texas, we demonstrate that the CIA policy consistently yields a significant and meaningful increase in audit profits as compared with the state's current audit selection policy and with five alternative active learning policies. Existing active learning policies may not only fail to enhance decisions but may even undermine them. We then apply CIA to improve direct targeting of individual donors based on the expected outcome of the solicitation. The application of CIA to the direct targeting task allows us to draw further insights about the implications of CIA for decision making and underscores the general value of our policy for improving a broad range of decisions, including health care fraud audits and targeted marketing.

2 Collaborative information acquisition

The use of active learning to improve sales tax audit decisions is a special case of a general problem that we define below. Consider a decision maker, such as a taxing authority, that must make a set of decisions D . These decisions provide benefits of interest, such as profits or the number of correct decisions, to which we will henceforth refer to

as utility. Each decision $d_i \in D$ is informed by a set M of two or more predictive models $M = \{M^1, M^2, \dots, M^m\}$.

A Collaborative Information Acquisition policy aims to acquire prospective training instances to augment the training data from which the models in M are induced so as to maximize the improvement per acquisition in the utility of decisions in D .

We consider sales tax audit decisions by a tax authority, where a decision about firm i , $d_i \in D$, corresponds to a choice of whether or not to initiate an audit. The decision-theoretic optimal choice is the one yielding the tax authority the largest expected utility. In particular, the expected utility from an audit of firm i is given by $EU = (p_i r_i) - Th$, where p_i is the probability of non-compliance, r_i denotes the revenue that can be recovered if the firm has not complied, and Th is a fixed value threshold, such as the cost of an audit. The expected revenue from an audit is given by $(p_i r_i)$, and is computed by inference from a *response* model, estimating the probability p_i , and from a *revenue* model, estimating the revenue that might be recovered. This decision structure corresponds to different scenarios for different values of Th . For example, the tax authority might initiate any profitable audit, that is, any audit in which expected revenue exceeds the cost of an audit Th . Alternatively, it might initiate only a subset of profitable audits in which expected revenue exceeds another possibly higher threshold Th .

In this paper, acquiring a training instance corresponds to obtaining the dependent variable values (labels) of prospective training instances for supervised learning. Note that the dependent variable values of prospective training examples for both the *response* and *revenue* models are acquired in an audit. Acquiring an audit yields information both about whether the company has complied with the law and about the revenue recovered in the audit.

Like existing policies, our information acquisition approach follows a sequential acquisition process. We consider an initial training set L of *labeled* instances, whose dependent variables are known, and a typically large set UL of *unlabeled* instances for which the dependent variable values can be acquired at a cost. At each acquisition phase, CIA evaluates the benefit of prospective acquisitions in UL , and then acquires a set of one or more learning instances that are added to the labeled training set L . After each acquisition phase, new models can be induced from L to inform the selection of additional training instances.

The benefit from a prospective acquisition may be estimated as the change the acquisition yields in expected utility relative to decisions informed by the *current* training data. This approach, however, disregards the sensitivity of induction to variations in the training sample. Future acquisitions will vary the training data and yield different predictive models to inform the decisions. Our CIA approach aims to take these variations into account to estimate the benefit of a prospective acquisition.

Our objective is to select acquisitions that will increase the likelihood of making an optimal decision, which requires correctly selecting the largest expected-utility action for decisions in D . We employ two heuristics to estimate the benefits of each acquisition. First, we assume that improvements from an acquisition are more likely if the acquisition pertains to decisions for which the current models are likely to yield a suboptimal choice. Second, we estimate the likelihood that a decision is optimal based on the distribution of decisions inferred across different data versions. Specifically, we approximate different instantiations of the training data by generating bootstrap samples from the available training data L , and induce a *response* and *revenue* model from each bootstrap sample T_j . For each pair of response and revenue models induced from data version T_j we then estimate the decision-theoretic optimal choice for each prospective acquisition i in UL , $\hat{d}_i^j$. Because the true optimal choice for decision i is not known, we capture the likelihood of inferring a suboptimal choice for

Algorithm 1 CIA: Collaborative Information Acquisition

Input: Set of prospective acquisitions UL each corresponding to a decision, an initial training set L , induction techniques $I_1, I_2, \dots, I_k$, batch size B , a stopping criteria

While (stopping criteria is not met) **Do**

1. Generate m bootstrap samples $T_1, T_2, \dots, T_m$, from L

2. **For** $j = 1$ **to** m **Do**

For $s = 1$ **to** k **Do**

 1) Apply I_s on T_j , resulting in model M_j^s

 2) Apply M_j^s to each instance i in UL

End For

End For

3. For each data sample T_j and instance i in UL : estimate the decision-theoretic choice $\hat{d}_i^j$

4. Acquire the set B of instances in UL for which the decisions $\hat{d}_i^j$, $j = 1, 2, \dots, m$, are most split

5. Remove B from UL , acquire B and add the acquired information to L

End While

Induce models $I_1, I_2, \dots, I_k$ from L to inform decisions

Output: model M_1 , model $M_2, \dots$, model M_k generated from L

decision d_i by the diversity in choices inferred across different data versions. In the extreme case in which the same choice is inferred from different versions, there is ample evidence in the data for this choice, and therefore we assume that this choice is optimal and will not benefit from additional acquisitions. A consistent choice across versions also suggests that acquiring additional information at a cost is unlikely to alter this choice and hence will not be cost-effective. By contrast, diverse choices across versions imply a higher likelihood that a suboptimal choice will be inferred, and we therefore assume that acquiring additional information is more likely to increase the likelihood of inferring the optimal choice. We capture the diversity in the decision-theoretic optimal choices estimated from different versions of the training data by the absolute difference between the frequencies of each choice. In the case of audit selection, the greatest diversity arises when half of the versions produce models which imply that auditing a given firm is desirable while the remaining versions recommend the opposite. A pseudo code of our collaborative information acquisition (CIA) policy is presented in Algorithm 1.

3 Empirical evaluation

Here we compare the outcomes of decisions informed with CIA with the outcomes of decisions informed by alternative approaches. The settings we consider here employ two primary phases: an information acquisition policy selects audits periodically to learn the models, and then the tax authority conducts a revenue-maximizing audit campaign informed by the learned models.

We employ data provided by Texas Comptroller of Public Accounts. The dataset reflects audits performed over a period of two years and is used by the state to produce the response and revenue predictive models. The data includes business information, such as taxpayer industry code (Standard Industrial Classification), business type (corporation, partnership, etc.), location, firm's capital, gross sales, information on employee and wages, information

Table 1 Tax audit data: descriptive statistics

No. non-compliant firms	2,931
No. compliant firms	2,069
Frequency of non-compliance	58.6 %
Median revenue, non-compliant firms (\$)	6,729
Average Revenue, non-compliant firms (\$)	25,884.9
Maximum Revenue (\$)	4,270,378
Minimum Revenue (\$)	38

on sales tax filings from recent years, information on other tax filings, and outcomes of prior audits. The data also includes the outcomes of sales tax audits carried out in the last campaign for each firm, that is, whether or not a firm has under-reported its earnings and any sales tax revenue recovered by the audit. Particularly informative features for inferring non-compliance include wages and a measure capturing the audit success rate of the relevant industry group. Particularly informative features for estimating audit revenue include the company's gross sales scaled by the average gross sales of peer firms and the ratio between reported taxable income and gross sales. Table 1 shows summary statistics of these data.

The models are induced from data reflecting only a fraction of the firms in the state. Each year only about 0.7 % of firms are audited. In addition, the state uses a subset of recent audits from a “moving window” to build the models; consequently, only about 3 % of the firms in the state are used for modeling. The available set of audits is not only small, but it also potentially lacks informative audits from the wide variety of firms in the state.

The taxing authority's objective is to conduct audits for which the expected revenue exceeds a given threshold amount. We first apply the CIA policy to identify informative audits from which the response and revenue models were induced and later use the learned models to identify desirable audits. Based on prior evidence of thresholds used in practice (Micci-Barreca and Ramachandran 2004), we consider two threshold values of Th , \$ 8,000 and \$ 10,000. These thresholds indicate respectively that 27 % and 23.44 % of all firms are desirable audits. In subsequent evaluations over a different domain, we show that our results are consistent across different settings.

Tax audit decision makers require models that are simple, and therefore the empirical evaluations that follow employ comprehensible predictive models used in practice to inform audit decisions. Specifically, we use J48 (Witten and Frank 1999), WEKA's implementation of C4.5 (Quinlan 1993) to induce the response model and we use WEKA's linear regression to estimate the revenue recovered in the case of non-compliance. In all the experiments we report here CIA employs 10 bootstrap samples at each acquisition phase.

Because our aim is to improve decisions, we evaluate alternative policies not by the accuracy of the models being deployed but by the profits from the decisions they inform. We report average results over 100 randomized experiments. In each experiment, we randomly partition the data into three exclusive subsets: an initial training set of 100 instances from which the initial set of models is induced, a pool of 1000 prospective companies whose audits can be acquired, and a test set of 1000 firms for evaluating audit decisions informed by each policy. To reduce experimental variance, the same partitions are used to evaluate all the policies. At each acquisition phase, we let each policy acquire 50 audits from the pool of prospective acquisitions. For each policy and after each acquisition phase we induce the response and revenue models from the augmented training data and subsequently apply the learned models to inform the decision-theoretic optimal choices for companies in the test set.

Table 2 Direct targeting data: descriptive statistics

Data set	20 % donors	40 % donors
Number of donors	1,002	2,048
Number of non-donors	4,019	3,083
Percent of donors	0.1996	0.3992
Maximum donation (\$)	200	200
Minimum donation (\$)	1	1
Avg donation among donors (\$)	15.83	15.6
Median donation among donors (\$)	13	13

In the figures that follow the initial performance shown for all policies is identical and reflects the decision performance enabled by the initial training set before any acquisitions are selected. New learning experiences are acquired by each policy until the entire pool of prospective acquisitions is exhausted. All policies also exhibit the same performance once the entire set of prospective acquisitions is acquired.

The training data available to the state (as well as to us) for supervised learning is not a representative sample, but constitutes firms that have been audited by the tax authority. However, we could not request that the state perform new company audits. This limits both the set of company audits that the active learning policies select from and the set of companies to which optimal decision making is applied. As such, our experiments reflect a scenario in which active learning is used to select or reject from a set of audits that has been pre-selected by the state. Nevertheless, an effective policy is likely to identify particularly informative audits in this setting. Similarly, the profit-maximizing audit decisions for our test set correspond to a scenario in which models learned with active learning are applied either to approve or to forgo audits proposed by the state. A comprehensive evaluation of CIA's benefits for audit selection can only be obtained once CIA is field tested by the tax authority, so that audits selected by CIA from the complete set of firms in the state can be acquired.

To compensate for this limitation in our primary data, we complement our empirical results with an evaluation of CIA's performance for similar decisions in a different domain where there are no constraints on the selection of informative instances or on the decisions made by the learned models. We consider the problem posed by the 1998 KDD-cup competition (Hettich and Bay 1999), in which the goal was to maximize the profit from a donation campaign by a charitable organization. The data consist of responses to solicitations by donors who had made their last donation between 13 and 24 months prior to the proposed campaign. To our knowledge the data reflects a representative sample of such donors. The models estimate the expected profit from a donor to allow the charity to target profitable donors in expectation. Table 2 shows descriptive statistics for this dataset. As in sales tax audit decisions, the decision-theoretic optimal choice is inferred from a *response* model and a *revenue* model. In the donation target application, the models estimate respectively the likelihood of response and the amount a prospective donor will contribute if a donation is made. For this evaluation we apply the same induction techniques as before. We also follow prior work for the choice of predictors and use only instances with actual donations to model the donation amount (Zadrozny and Elkan 2001). We employ CIA to acquire the donor responses from which the predictive models are learned in order to improve subsequent targeting decisions. We also aim to explore how different policies perform as the decision parameters change. We create several versions of the problem by varying the proportion of donors and the risk (loss) from a futile solicitation. Furthermore, the solicitation task defined by the 1998 KDD cup competition entails that the trivial solution of soliciting

all customers yields performance that is only marginally improved with better estimations of the decision-theoretic choices. To capture differences among alternative policies, we evaluate the performances of the policies when 20 % and 40 % of solicitations yield a donation, and vary the risk (loss) incurred from futile solicitations. We consider increasingly higher solicitation costs of \$ 1, \$ 3, and \$ 5 with 20 % donors, and an additional set of higher costs of \$ 7, \$ 9, and \$ 11 with a higher proportion of 40 % donors.

3.1 Alternative information acquisition policies

We compare the benefits from decisions informed by CIA with those obtained by five alternative acquisition policies and a uniform policy that acquires learning experiences uniformly at random.

Three alternative acquisition policies are accuracy-centric and aim to improve the predictive accuracy of one of the models informing the decisions. We apply two of these policies, Uncertainty Sampling (US) (Lewis and Gale 1994) and Query-by-Bagging (QBB) (Abe and Mamitsuka 1998), to improve the response model's accuracy, and the third policy, proposed by Krogh and Vedelsby (1995) (henceforth, KV), to improve the revenue model's accuracy. Our fourth alternative acquisition policy, Alternate Selection (hereafter referred to as Alternating) is a type of Multi-Task Active Learning (MTAL) (Reichart et al. 2008) that aims to simultaneously improve the predictive accuracies of *multiple classification* models. Alternating and the other MTAL policy, Rank Combination, exhibit comparable performance (Reichart et al. 2008). However, the accuracy of each of the two MTAL policies is comparable to or lower than that achieved by a single-model accuracy-centric policy applied to each of the models. In our empirical evaluation we employ a variant of the Alternating policy. The original Alternating policy aimed to improve two classification models by alternating between acquisitions of informative learning experiences for each of the models. We adapt this policy to alternate between acquisitions of particularly informative instances for the response and revenue models. We select informative instances for the response model using the Uncertainty Sampling active learner (Lewis and Gale 1994) and for the revenue model using the policy proposed by Krogh and Vedelsby (1995). Note that an MTAL policy implicitly assumes that all improvements in any of the models are equally desirable, regardless of their implications for decisions. The objective of our comparisons is to evaluate whether CIA can improve decisions through its capacity to prefer improvements in one of the models over the other.

Our final alternative policy, GOAL, was the first decision-centric active learner (Saar-Tszechansky and Provost 2007). GOAL pursues acquisitions to improve a classification model's probability estimation if these improvements are also likely to yield better decisions. However, unlike CIA, GOAL does not consider the implications of prospective acquisitions for other models informing decisions. Our evaluation explores whether CIA's consideration of all the models yields better decisions than enabled by GOAL.

Finally, note that all active learners bias the training set to achieve a given objective. When multiple models are induced from the data, a bias that improves one model may undermine the other with potentially adverse outcomes for decision making. By contrast, a uniform acquisition policy is bias-free, as it is equally beneficial for acquiring learning examples for any model. We aim to establish whether different active learners consistently yield decisions that are at least comparable to decisions informed by a representative sample drawn uniformly at random.

Fig. 1 Sales tax audit profits obtained with CIA relative to alternative policies, with audit cost at \$ 8000 per firm

4 Results

Our empirical evaluations have two general objectives. First, we aim to establish whether the bias introduced by each active learner consistently yields comparable or better decisions than those informed by the Uniform policy. We posit that a bias that does not align with the decision-making objective or does not accommodate learning of multiple models may not improve or may undermine the decisions the models inform. Second, we discuss the benefits of CIA relative to each alternative.

For tax audit decisions, Fig. 1 shows the sales tax profits obtained per decision after each acquisition phase by CIA and each of the alternative active policies. Note that profits are computed per decision made, including decisions in which an audit is not initiated. Similarly, Fig. 2 presents profits per decision for the direct targeting problem. Following prior work (Melville et al. 2004), Table 3 summarizes the average percent change in profits obtained through CIA compared with that obtained through the alternative policies across all acquisition phases. For each pair-wise comparison, differences are marked in boldface if they are statistically significant according to a paired t -test ($p \leq 0.01$).

Fig. 2 Profits from direct targeting obtained with CIA and alternative policies

Overall, we find that audits acquired by CIA-produced models yield significantly higher profits from effective audit decisions as compared with profits produced by the Uniform policy or any of the alternative active learners.

The magnitude of CIA's advantage over its alternatives is substantial. Across all acquisition phases, CIA increases sales tax profits by an *average* of 4 % or more over the level obtained by the Uniform policy. In the three consecutive acquisition phases in which CIA produces the greatest improvement over the current policy, CIA yields an average 6.7 % increase in profits. Note that the average increase over all acquisition phases is a conservative estimate of CIA's benefits because it considers improvements in the initial and final acquisition phases where the training data available to different policies are practically identical. Also note that because the Uniform policy selects audits uniformly at random from a pool of firms selected for audit by the state, our results demonstrate that CIA constitutes a reliable, advantageous alternative to the existing paradigm. The state taxing authority reports \$ 90 million in annual tax revenue from audits. Our projections suggest that if CIA is used to

Table 3 Percent increase in profit per decision obtained by CIA relative to alternatives. Tax refers to Tax Audit, Targeting I and Targeting II refer to Direct Targeting with 20 % and 40 % donors, respectively

Domain	Uniform	GOAL	Alternating	US	QBB	KV
Tax, cost \$ 8,000	4.5	4.4	4	4.4	3.4	3.7
Tax, cost \$ 10,000	4	4.5	3.69	4	3.1	3.2
Targeting I, cost \$ 1	1.5*	−0.1	1.3	1.8	3.6	1.3
Targeting I, cost \$ 3	9.2	1.1	9.4	10.4	9.3	9.3
Targeting I, cost \$ 5	13.3	80.6	6.7	19.2	1	7.3*
Targeting II, cost \$ 1	0.5	1	0.7	1.4	4.9	0.7
Targeting II, cost \$ 3	2.1	3	2.4	4.4	4.6	2.4
Targeting II, cost \$ 5	9.9	7.1	10.8	14.7	18.8	9
Targeting II, cost \$ 7	9.1	10	12.3	17.9	21	10.8
Targeting II, cost \$ 9	17.3	16.7	28.2	39.7	25.8	25.3
Targeting II, cost \$ 11	27.6	24	70	59.3	27.7	59

Results in bold are statistically significant at $p \leq 0.01$.

*Statistically significant at $p \leq 0.05$.

select only from audits already approved by the state, it can potentially increase the state's revenue by \$ 3.6–6 million dollars annually. The benefits of CIA are likely to be even greater outside the US, where tax compliance rates are generally below the US level (Christie and Holzner 2006).

The benefits of CIA for targeting decisions, where CIA acquires learning experiences from the natural distribution of donors, confirm CIA's observed benefits for sales tax audit decisions. Together, our results for sales tax audit and direct targeting show that regardless of the pool of prospective learning experiences available for acquisition, CIA consistently identifies informative learning experiences that improve sales tax audit and direct targeting decisions better than acquisitions from the Uniform policy.

A comparison between CIA and each of the alternative policies for the tax audit and direct targeting domains shows that CIA's choice of learning experiences produces higher profits in an overwhelming majority of decisions. Among sixty-six pairwise comparisons for different decision settings and alternative policies, CIA produces statistically significantly higher profits in sixty-three tasks and comparable profits in three tasks.

Our empirical evaluations also suggest that existing policies may perform worse even than the Uniform policy. For the targeting campaign, as shown in Fig. 2, all alternative policies yield worse decisions than those informed by representative acquisitions in some settings. QBB, for example, exhibits highly variable performance. QBB produces excellent performance in some settings but yields substantially lower profits than the Uniform policy in other settings (see Fig. 2). In contrast, CIA's improvement in decisions over the uniform benchmark is consistent and is unmatched by any alternative policy. We now highlight CIA's benefits relative to each of the alternative policies. We find that GOAL often produces lower revenues than those generated by CIA, and it is comparable to CIA in only two of the settings. While GOAL's acquisitions aim to improve decisions, GOAL only considers the response model and implicitly assumes that the revenue model is unaffected by acquisitions. Our results demonstrate that because this assumption is violated in our setting, GOAL does not yield benefits consistently comparable to those exhibited by CIA.

Table 4 Percent increase in profit per audit decision obtained by CIA relative to alternatives using *Bagging* for response modeling

Domain	Uniform	GOAL	Alternating	US	QBB	KV
Tax, cost \$ 8,000	4.86	4.64	4.02	1.87	1.03	3.81
Tax, cost \$ 10,000	6.61	7.11	5.29	2.87	2.36	5.14

Results in bold are statistically significant at $p \leq 0.01$.

The bias introduced by the accuracy-centric policies QBB and US aims to improve the response model's generalization accuracy. QBB can achieve good decision performance that is comparable to that of CIA. However, as shown in Fig. 2, QBB also pursues learning experiences that are not cost effective for future decisions, and can yield substantially worse decisions than achieved by the uniform policy. Uncertainly Sampling and KV's choice of learning experiences consistently yields worse decisions than those informed by CIA.

Our results for GOAL and the accuracy-centric active learners highlight the potential drawback of a policy that considers only one of the models informing the decisions. However, while the Alternating approach improves both models, it still fails to yield substantial benefits. Therefore, the advantages shown by CIA relative to Alternating can be attributed to forgoing acquisitions that improve either one of the models or both if these improvements are not likely to improve the decisions.

Note that if a policy effectively avoids uninformative or even harmful acquisitions, the consequences of these acquisitions may be revealed in later batches, as the policy exhausts the pool of prospective acquisitions (UL). CIA aims to avoid acquisitions of firms for which estimated expected revenues are consistently well below or above the relevant threshold. Such instances may also be outliers that can undermine modeling revenue or the probability of non-compliance. As shown in our empirical results, CIA's profits from audit decisions decline as the pool of audits is exhausted and the remaining audits are added to the training data. Thus, the acquisitions avoided by CIA early on are not only uninformative, but also undermine decisions.

The advantages of CIA appear to be robust to changes in modeling techniques. We performed additional evaluations of CIA's benefits for sales tax audit selection using Bagging (with J48 as base model) instead of J48 to estimate the likelihood of response. Because Bagging is a popular and effective modeling technique, this evaluation can shed light on CIA's benefits for audit selection when different modeling techniques are being used to inform audit decisions. As shown in Table 4, our results confirm that CIA performs well using Bagging as well: CIA exhibits consistent and significant benefits over Uniform and each of the alternative active policies. The average increase in audits profits relative to the uniform policy ranges between 4.86 % to 6.61 %.

4.1 Insights into CIA's selective acquisitions

A key motivation for collaborative information acquisitions is that improvements in the predictive accuracies of the models do not always yield better decisions. As it acquires more instances, CIA improves the underlying models' predictions. However, CIA achieves its economies through a selective improvements in models: it forgoes improvements if they will not likely change the decisions the models currently infer. CIA's acquisitions produce better decisions, but the *average* accuracy of any of the models induced from these acquisitions may be equivalent, better, or even worse than the average accuracy that can be

Fig. 3 Profits are higher with CIA while the models' average error can be larger

achieved with a representative sample. Figure 3 demonstrates this outcome in two settings. CIA yields better decisions, while the average error of CIA's revenue model is higher than or comparable to that of a model induced from a representative sample.

Finally, we examine the overlap between audits acquired by CIA and by each of the alternative policies. We compute the overlap obtained after 5 acquisition phases, after which a sufficient number of examples have been acquired to allow the training set compiled by the different policies to diverge. Table 5 shows percent overlap with each policy, averaged over 10 experiments. As shown, a majority of the audits acquired by CIA are not acquired by the alternative policies. Specifically, only 30 % to 35 % of the examples acquired by CIA are the same as those acquired by the alternative active policies. An estimation of the overlap for some of the donor targeting problem settings revealed a similar pattern. Improvement in profitable targeting is difficult in the donor targeting problem, but we did not observe the overlap consistently decrease or increase as a result. Specifically, for a solicitation cost of \$ 1, 24–38 % of CIA's acquisitions were the same as those acquired by alternative policies.

5 Related work

To our knowledge, no prior work has explored collaborative active learning of multiple predictive modeling tasks to improve cost-effectively the decisions the models inform. Existing policies either address a single predictive modeling task (e.g., Abe and Mamitsuka 1998; Krogh and Vedelsby 1995; Lewis and Gale 1994; Roy and McCallum 2001; Saar-Tsechansky and Provost 2007), or aim to improve the predictive accuracy of multiple models (Reichart et al. 2008).

Table 5 Percent overlap in acquisitions acquired by CIA and by alternatives

Domain	Uniform	GOAL	Alternating	US	QBB	KV
Tax, cost \$ 8,000	25.36	29.36	34.88	31.80	33.80	37.16
Tax, cost \$ 10,000	24.00	30.40	33.44	31.00	32.48	35.04
Targeting I, cost \$ 1	24.48	38.64	32.20	32.40	24.08	32.32
Targeting I, cost \$ 5	25.16	14.92	33.12	32.44	31.24	32.56
Targeting II, cost \$ 1	26.20	27.04	35.16	27.88	28.76	37.28
Targeting II, cost \$ 9	24.60	28.00	30.92	28.76	27.76	32.08

Most work on active learning considers accuracy-centric policies for improving a *single* model, typically a classification model, but sometimes a regression model (e.g., Abe and Mamitsuka 1998; Krogh and Vedelsby 1995; Lewis and Gale 1994; Roy and McCallum 2001). Settles (2009) offers an extensive review of active learning for classification models, and Attenberg et al. (2011) include in-depth discussion of important practical objectives for such policies. Importantly, when decisions are informed by multiple models, such policies do not consider the implications of prospective acquisitions on other models and on any decisions the models might inform. Our empirical evaluations show that this limitation yields decisions that can be worse than can be achieved with a uniform policy.

Reichart et al. (2008) have proposed Multi-Task Active Learning (MTAL): accuracy-centric active learning for multiple classification models (Reichart et al. 2008). However, MTAL implicitly assumes that improvement in the accuracy of any of the models is equally desirable and does not consider any decision tasks that the models might inform.

Some active learning policies for a single classification model consider a cost-sensitive setting (e.g., Kapoor et al. 2007; King et al. 2004). However, these policies do not address either settings in which the benefits (e.g., from successful audits) *differ* across instances or choices that require inferences from multiple predictive models of different kinds.

Prior work has shown that costly improvements in predictive accuracy do not always result in better decisions and has proposed the first *Decision-centric* goal-oriented active learner (GOAL) (Saar-Tsechansky and Provost 2007). GOAL pursues acquisitions to improve selectively a single classification model's probability estimation when these improvements are likely to yield better decisions. However, for the settings we explore here, GOAL does not consider the implications of prospective acquisitions on the regression model's estimations.

The CIA framework we propose here might be extended in the future to consider problems where other types of information can be acquired, such as feature values (e.g., Lizotte et al. 2003; Saar-Tsechansky et al. 2009), and settings where the acquired values may be incorrect (Sheng et al. 2008). Thus far, such policies have considered a single classification model with the objective of improving the model's predictive accuracy.

6 Fielding CIA

Over the past decade a growing number of tax authorities in the US, Canada, and Europe have begun using predictive modeling to inform tax audit decisions. Resource constraints have often led authorities to limit exploratory audits that might improve future decisions and instead to focus on myopic optimization of revenues. CIA offers an opportunity to begin the process of improvement by selectively identifying particularly informative audits

that can cost-effectively improve future audit revenues. The work we report here was done in partnership with Elite Analytics, a consulting firm with extensive experience with audit selection and non-compliance projects. Elite Analytics has been consulting with tax authorities in North America to help them use predictive modeling effectively with the objectives of increasing tax revenue and improving compliance.

In this paper we report results for a tax authority whose objective is to conduct audits for which the expected revenue exceeds a given threshold amount. The threshold amount varies depending on the auditing resources available for any given campaign. To calculate the expected revenue for a given firm, the state's audit decisions are informed by a revenue and response predictive models that estimate the audit revenue and the likelihood of non-compliance, respectively.

Before existing practices should be changed, CIA should be partially deployed for evaluation to assess its full impact. Towards that end, Elite Analytics has shared the revenue curves produced by CIA with one of the tax authorities, and we are pursuing approval of CIA for partial deployment in high-volume, *desk-based* audits. Since such audits are limited in scope and do not involve auditors' visit to a company's offices, they are relatively fast to complete, and thereby expedite the evaluation of the policy. Evidence from the successful partial deployment of CIA for desk audits is anticipated to pave the way to full deployment, as well as to facilitate deployment in other audit programs and across different tax authorities.

From an operational perspective, integrating CIA within an audit program in which predictive models are already used is straightforward because CIA merely uses inferences from these models to identify particularly informative audits. The predictive models currently used can be applied to produce estimations for prospective audits, and then CIA can use these estimations to prioritize and select particularly informative audits.

Finally, CIA can readily be used to improve similar decisions in different domains. We hope that this work will promote the adoption of CIA to benefit other important tasks.

Commentary by Dr. Daniele Micci-Barreca and Dr. Satheesh Ramachandran, Elite Analytics Sales tax remains the single most important revenue source for state governments in the US, and similarly, sales or value added tax (VAT) represent a significant source of revenue for many governments around the world. Auditing tax returns is the principal means used by tax administrations to maintain, enforce and encourage compliance. Since auditing resources are limited, cost-effective management of auditing programs is a significant step towards improving tax compliance, increase revenue and ultimately reduce the "tax gap", while maintaining fairness in the sharing of the tax burden across the population. A key drawback of the existing approaches is that they produce myopically optimal revenues, whereby missing the scope to explore potentially informative audits that will improve profits or may even discover unknown areas of non-compliance. As explained by the authors, the possible implications of selective acquisitions through active learning to enhance audit campaigns is not well understood, and, to our knowledge, has not been explored in practice. This study and its results are innovative and forward looking, with meaningful potential implications. The method presented here will enable a new paradigm for audit selection, where explorations are decided based on their contribution to future audit decisions, and taking into account all the models involved. The resource constraints that tax agencies face, have thus far limited their capabilities to invest resources to improve their long-term analytical capabilities. However, this work offers a cost-effective policy to do so, focusing on the agency's objective to increase future revenues. Based upon our experience, possible practical challenges in incorporating active learning methods in audit agencies pertain to the perceived risk in

changing existing decision making practices, particularly if the change involves data-driven algorithms. Given that field audits take a significant amount of time to complete, our recommendation to expedite adoption is to deploy CIA first with high-volume audit programs, particularly desk-based audits program. As opposed to field audit programs, desk audits are relatively faster, allowing the generation of informative training examples via CIA within a shorter time frame. Fielding CIA will also involve a champion-challenger framework, whereas a subset of the audits will be selected via CIA. We believe this framework will expedite the path to adoption, allowing the agency to capture the benefits of CIA vis-a-vis the legacy audit selection approaches in a short timeframe and with fewer resources.

7 Conclusions and lessons for the machine learning community

This research has been motivated by the sales tax audit selection problem in which decisions are informed by multiple predictive models of different kinds. Efficient audit selection and debt collection (Abe et al. 2010) are vital steps toward decreasing the staggering losses from non-compliance in the United States and other countries. However, the work we present here also demonstrates the growing importance for research of considering the context in which machine learning is applied. In our work, a careful examination of the task's objective was instrumental in revealing the shortcomings of existing approaches, giving rise to a new information acquisition problem. Our research has also revealed the risks of a naive application of prior policies to inform decisions. We show that improvements in predictive accuracy are not always justified when model accuracy is not the end goal. Furthermore, we show that the bias introduced by single-model active learning policies can be detrimental to the task.

Because decision-centric acquisitions are tailored to improve certain repetitive decisions, CIA will not be the policy of choice if the objective is to improve the models' average accuracies. Similarly, CIA will not be the best policy to address problems in which the same set of predictive models inform *different* kinds of decisions. Traditional, accuracy-centric active learning policies are best suited to these cases.

Machine learning research can extend its impact on practice by considering the overall objectives of systems of which machine learning techniques are components. The sales tax revenue problem corresponds to one class of decisions. However, the general CIA problem arises in different classes of decisions as well. There are therefore opportunities for future work to develop CIA policies for decision tasks in addition to those we consider here. We hope that this research will motivate future work to explore these opportunities.

Acknowledgements We thank Dr. Daniele Micci-Barreca and Dr. Satheesh Ramachandran for their introduction to the tax audit problem and for their generous collaboration. We also thank anonymous Machine Learning journal reviewers and the editors of this special issue for their insights and for thoughtful comments and suggestions that have improved the paper.

References

- Abe, N., & Mamitsuka, H. (1998). Query learning strategies using boosting and bagging. In *Proceedings of the fifteenth international conference on machine learning* (pp. 1–9). San Mateo: Morgan Kaufmann
- Abe, N., Melville, P., Pendus, C., Reddy, C. K., Jensen, D. L., Thomas, V. P., Bennett, J. J., Anderson, G. F., Cooley, B. R., Kowalczyk, M., Domick, M., & Gardinier, T. (2010). Optimizing debt collections using constrained reinforcement learning. In *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '10)* (pp. 75–84). New York: ACM.

- Attenberg, J., Melville, P., Provost, F., & Saar-Tsechansky, M. (2011). Selective data acquisition for machine learning. In *Cost-sensitive machine learning*. Boca Raton: CRC Press.
- Christie, E., & Holzner, M. (2006). *What explains tax evasion? An empirical assessment based on European data*. wiiw working papers 40, The Vienna Institute for International Economic Studies, wiiw.
- Cohn, D., Atlas, L., & Ladner, R. (1994). Improving generalization with active learning. *Machine Learning*, 15(2), 201–221.
- Cornelius, T. (2012). Wisconsin budget basics guide, Wisconsin budget project. <http://www.wisconsinbudgetproject.org/wp-content/uploads/2012/07/budgetbasics.pdf>. Accessed 2013-05-11.
- EOI (2012). Washington state budget 101. http://www.eoionline.org/tax_reform/fact_sheets/WashingtonStateBudget101-Nov12.pdf. Accessed 2013-05-11.
- Hettich, S., & Bay, S. D. (1999). In *The UCI KDD archive*.
- IRS (2012). IRS releases new tax gap estimates; compliance rates remain statistically unchanged from previous study. <http://www.irs.gov/uac/IRS-Releases-New-Tax-Gap-Estimates;-Compliance-Rates-Remain-Statistically-Unchanged-From-Previous-Study>. Accessed 2013-05-15.
- Kapoor, A., Horvitz, E., & Basu, S. (2007). Selective supervision: guiding supervised learning with decision-theoretic active learning. In *International joint conference on artificial intelligence*.
- King, R. D., Whelan, K. E., Jones, F. M., Reiser, P. G. K., Bryant, C. H., Muggleton, S. H., Kell, D. B., & Oliver, S. G. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427, 247–252.
- Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. In *Advances in neural information processing systems* (pp. 231–238). Cambridge: MIT Press.
- Lewis, D., & Gale, W. A. (1994). A sequential algorithm for training text classifiers. In *Proceedings of the 6th international conference on knowledge discovery and data mining*.
- Lizotte, D., Madani, O., & Greiner, R. (2003). Budgeted learning of naive-Bayes classifiers. In *Proceedings of the 19th conference on knowledge uncertainty in artificial intelligence*.
- Melville, P., Saar-Tsechansky, M., Provost, F., & Mooney, R. (2004). Active feature-value acquisition for classifier induction. In *Proceedings of the 3rd international conference on data mining*.
- Micci-Barreca, D., & Ramachandran, S. (2004). *Improving tax administration with data mining*. White paper. Elite Analytics LLC.
- Quinlan, J. R. (1993). *C4.5: programs for machine learning*. San Mateo: Morgan Kaufmann.
- Reichart, R., Tomanek, K., & Hahn, U. (2008). Multi-task active learning for linguistic annotations. In *ACL*.
- Roy, N., & McCallum, A. (2001). Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the 18th international conference on machine learning* (pp. 441–448).
- Saar-Tsechansky, M., & Provost, F. (2007). Decision-centric active learning of binary-outcome models. *Information Systems Research*, 18(1), 1–19.
- Saar-Tsechansky, M., Melville, P., & Provost, F. J. (2009). Active feature-value acquisition. *Management Science*, 55(4), 664–684.
- Settles, B. (2009). *Active learning literature survey*. Tech. rep., University of Wisconsin-Madison.
- Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008). Get another label? Improving data quality and data mining using multiple, noisy labelers. In *KDD* (pp. 614–622).
- Witten, I. H., & Frank, E. (1999). *Data mining: practical machine learning tools and techniques with Java implementations*. San Mateo: Morgan Kaufmann.
- Zadrozny, B., & Elkan, C. (2001). Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of the 7th international conference on knowledge discovery and data mining* (pp. 204–212). New York: ACM.