On the Tractability of Rule Discovery from Distributed Data

Martin Scholz
Artificial Intelligence Group
Department of Computer Science
University of Dortmund, Germany
scholz@ls8.cs.uni-dortmund.de

Abstract

This paper analyses the tractability of rule selection for supervised learning in distributed scenarios. The selection of rules is usually guided by a utility measure such as predictive accuracy or weighted relative accuracy. A common strategy to tackle rule selection from distributed data is to evaluate rules locally on each dataset. While this works well for homogeneously distributed data, this work proves limitations of this strategy if distributions are allowed to deviate. The identification of those subsets for which local and global distributions deviate, poses a learning task of its own, which is shown to be at least as complex as discovering the globally best rules from local data.

1. Introduction

The induction of interesting rules from classified examples has been studied extensively in the Machine Learning literature throughout the last decades. A variety of metrics like predictive accuracy, precision, or the binomial test function have been suggested to formalise the notions of interestingness and usefulness of rules. [4] gives an overview of different metrics and illustrates the differences by means of ROC isometrics. There are several learning tasks that are formulated as optimisation problems with respect to a specific metric. Classifier induction and subgroup discovery are two examples. Usually it is assumed that all the available data is accessible to a single learner. In this case the metrics allow to identify a set of patterns that maximise the selected utility function. The amount of data necessary to identify the best rules with high probability depends on the evaluation metric [7], and can be considered as an indicator of complexity from an information theoretic point of view.

There are several learning scenarios in which the access to the available data is restricted. In the domain of knowledge discovery in databases, for example, the data is often split to different sites and may not be communicated at the level of single examples. Among the reasons are privacy issues and costs.

Learning tasks can be adopted to distributed scenarios in various ways. The objective of this work is to analyse the corresponding increase in complexity of rule selection, compared to non-distributed learning. Due to its generality the task of subgroup discovery fits nicely into this framework. It allows to specify the utility function used for pattern selection as a parameter [5]. Each subgroup is usually represented by a Horn logic rule, so utility functions are specific kinds of rule selection metrics. This paper investigates in which situations a local evaluation of rules may help to identify globally best rules, and how corresponding learning tasks are related to each other.

2. Standard subgroup discovery

This sections discusses the task of non-distributed subgroup discovery. Given is a set of m classified examples $\mathcal{E} := \langle x_1, y_1 \rangle, \ldots, \langle x_m, y_m \rangle$ from $\mathcal{X} \times \mathcal{Y}$. $\mathcal{X}$ defines an instance space and $\mathcal{Y}$ a set of labels. The representation language ($\mathcal{H}$) contains logical rules, denoted as $A \rightarrow C$. Antecedents A are identified with their corresponding subsets of $\mathcal{X}$, while conclusions C predict a label from $\mathcal{Y}$.

Rules are evaluated with respect to a distribution function over $\mathcal{X}$. This work confines itself to descriptive learning, so given a single database or example set $\mathcal{E}$ it is often appropriate to assume a uniform distribution D over $\mathcal{E}$.

Definition 1 *The* coverage (Cov) *of a rule $A \rightarrow C$ under distribution D is defined as the probability that it is applicable for an example $\langle x, y \rangle$ sampled $\sim D$:*

$$\textsc{Cov}_D(A \rightarrow C) := Pr_{\langle x,y \rangle \sim D}\left[x \in A\right]$$

Definition 2 *The* bias *of a rule $A \rightarrow C$, $C \in \mathcal{Y}$ under D is defined as the difference between the conditional probability of C given A and the default probability of C:*

$$\begin{aligned} \textsc{Bias}_D(A \rightarrow C) \quad &:= \quad Pr_{\langle x,y \rangle \sim D}\left[y = C \mid x \in A\right] \\ &- \quad Pr_{\langle x,y \rangle \sim D}\left[y = C\right] \end{aligned}$$

Def. (1) and (2) allow to state a broad class of metrics.

Definition 3 *Functions $f : \mathcal{H} \times D \to \mathbb{R}$ satisfying the following constraint for all r, r' are called* utility functions*:*

$$(\textsc{Cov}_D(r) \geq \textsc{Cov}_D(r')) \wedge (\textsc{Bias}_D(r) \geq \textsc{Bias}_D(r') > 0) \\ \Rightarrow f(r, D) \geq f(r', D)$$

Additionally, if one of the inequalities is strict, then $f(r, D) > f(r', D)$.

The most commonly used class of utility functions for subgroup discovery [5] is given by the following definition:

Definition 4 *For a given parameter α and distribution D the utility (or quality) $\underline{\mathrm{Q}}_D^{(\alpha)}$ of a rule $r \in \mathcal{H}$ is defined as*

$$\mathrm{Q}_D^{(\alpha)}(r) := \textsc{Cov}_D(r)^{\alpha} \cdot \textsc{Bias}_D(r).$$

The parameter α allows for a data- and task-dependent trade-off between coverage and bias. Def. 4 covers metrics that are factor-equivalent to the binomial test function ($\alpha = 0.5$), weighted relative accuracy ($\alpha = 1$), and a function applied to put higher emphasis on coverage ($\alpha = 2$).

Def. 3 is broad enough to also cover predictive accuracy, which is equivalent to $\mathrm{Q}_D^{(1)}$ for binary prediction tasks with equal default probabilities for both classes, and which is still monotone in Cov and Bias, otherwise. The similarity between rule selection metrics for different skew ratios is discussed in [3].

In association rule mining [1] rules are filtered (or pruned) by their support (Cov) and confidence. The latter is monotone in the Bias, although the default probability is usually ignored. When support and confidence are combined (respecting monotonicity) to find a ranking of most interesting rules, this problem can also be considered as a specific case of subgroup discovery.

3. Homogeneously distributed data

A first extension towards distributed subgroup discovery is to assume that several sets of data are available, which all obey a common underlying probability distribution. One can think of the different sets as generated by bootstrapping from a single, global dataset. In such a case local and global subgroups are basically identical. However, due to statistical deviations caused by bootstrapping and the smaller size of example sets, some of the rules with lower global utilities might be found among the n best subgroups evaluated locally at each site.

Choosing $\mathrm{Q}^{(1)}$ (Def. 4), the probability that the utility function deviates locally from the true (global) value by more than a fixed constant $\epsilon \in \mathbb{R}^+$ can be bounded by Chernoff's inequality. This probability decreases exponentially with a growing number of examples. Sample bounds have been proven for different utility functions [7], especially for $\mathrm{Q}^{(\alpha)}$ with $\alpha \in \{.5, 1, 2\}$. Accordingly, the n-best subgroups problem has been adopted to a probabilistic scenario, in which utility functions are evaluated using i.i.d. samples [7]:

Definition 5 *Let $\delta \in (0, 1)$ denote a given minimum confidence and $\epsilon \in \mathbb{R}^+$ denote a given maximal error. Then the* approximate n-best hypotheses problem *is to identify a set G of n hypotheses from a hypothesis space $\mathcal{H}$, such that with confidence $1 - \delta$*

$$(\forall h' \in \mathcal{H} \setminus G) : \mathrm{Q}(h') \leq \min_{g \in G} (\mathrm{Q}(g) + \epsilon)$$

The results reported for this problem directly apply to homogeneously distributed datasets: For large local datasets the probability of missing a subgroup that is globally much better than the locally best ones is reasonably small.

4. Inhomogeneously distributed data

This section addresses the situation in which data is split to different sites, but no distributional assumption can be made. First of all the notation for different databases is introduced. The example set $\mathcal{E}$ is composed of k subsets $\mathcal{E}_1, \ldots, \mathcal{E}_k$ that were sampled from different probability distributions. Let D_i denote the distribution at site i for the corresponding example set $\mathcal{E}_i \subseteq \mathcal{E}$, and let D denote the global distribution over $\mathcal{E}$. D is a weighted average of the local distributions.

Local Cov and Bias of a rule $A \to C$ at site i can be expressed in terms of Def. 1 and 2, replacing D by D_i, e.g.

$$\begin{aligned} \textsc{Bias}_{D_i}(A \to C) \quad &:= \quad Pr_{\langle x,y \rangle \sim D_i}[y = C \mid x \in A] \\ &- \quad Pr_{\langle x,y \rangle \sim D_i}[y = C] \end{aligned}$$

refers to the local Bias at site i. Accordingly, a local utility function evaluates each rule $A \to C$ by

$$\mathrm{Q}_{D_i}^{(\alpha)}(A \to C) = [\textsc{Cov}_{D_i}(A \to C)]^{\alpha} \cdot \textsc{Bias}_{D_i}(A \to C).$$

The first task stated in this setting is to find subgroups that globally perform well, given a discovery procedure that evaluates rules locally. The idea is, that if one of the globally best rules appears poor at any site, then it obviously needs to perform even better at some other. For this reason one could expect that the globally best rules are easily found at the local sites, even if the local distributions differ. A similar property eases frequent itemsets mining from distributed data [2].

In the case of homogeneously distributed data as discussed in Sec. 3, the marginal distributions over $\mathcal{X}$ and the conditional probabilities of the target given $x \in \mathcal{X}$ were identical at all sites. In order to quantify by how much each of these assumptions is weakened the following definitions are useful.

Definition 6 *Two distributions* $D_1, D_2 : \mathcal{X} \to \mathrm{I\!R}^+$ *are called* factor-similar up to γ *for an* $A \subset \mathcal{X}$ *and* $\gamma > 1$*, if*

$$(\forall x \in A) : \gamma^{-1} \leq \frac{D_i(x)}{D(x)} \leq \gamma.$$

Definition 7 *For an* $A \subseteq X$ *two joint distributions* $D_1, D_2 : \mathcal{X} \times \mathcal{Y} \to \mathrm{I\!R}^+$ *are called* conditionally similar up to ϵ, $\epsilon > 0$*, if*

$$(\forall \langle x, y\rangle \in A \times \mathcal{Y}) : \left| \frac{D_1(x,y)}{D_1(x)} - \frac{D_2(x,y)}{D_2(x)} \right| \leq \epsilon.$$

Please recall that utility functions are defined based on distributions underlying the example sets. For this reason Def. 6 and 7 do not require the same set of examples to be observable at all sites to allow for finite bounds.

The following theorem shows, that if the assumption of homogeneously distributed data made in Sec. 3 is weakened at all, then it is possible to obtain drastically different sets of best rules when evaluating a quality function globally and locally.

Theorem 1 *Let* G_i *denote the set of* n *best rules for each site* $i \in \{1, \ldots k\}$ $(k \geq 2)$*, given an arbitrary utility function. Let* G *denote the set of* n *best rules with respect to the global distribution. Then it is possible in the general case, that every* $x \in \mathcal{X}$ *is covered by at most one ruleset from* $\{G, G_1, \ldots, G_k\}$*, where a ruleset is said to cover* x *if one of its elements does. This statement even holds in the following two cases:*

1. *For all local sites* $i \in \{1, \ldots, k\}$ *the conditional distributions of* $\mathcal{X} \times \mathcal{Y}$ *are identical, and each local marginal distribution of* $\mathcal{X}$ *is factor-similar to the global one up to an arbitrarily small* $\gamma > 1$ *for any subset of* $\mathcal{X}$*.*
2. *The global and local marginal distributions of* $\mathcal{X}$ *are equivalent, and global and local joint distributions of* $\mathcal{X} \times \mathcal{Y}$ *are conditionally similar up to an arbitrarily small* $\epsilon > 0$*.*

A proof is given in an extended version of this article [8]. Theorem 1 implies that rules globally performing best are not necessarily among the n locally best rules at *any* site. Even for arbitrarily unskewed data, formalised in terms of Def. 6 and 7, the best rules collected from all sites, including the globally best rules, may be completely disjoint, in the sense that no example is covered twice. Please note that unlike for the case of homogeneously distributed data this is not a problem of misestimation. Theorem 1 applies to arbitrarily large sample sizes.

Although finding the globally best rules from local data is not possible in the worst case, finding approximately best rules might still be tractable. The following theorem gives a tight bound on the difference between locally and globally evaluated utility, for simplicity assuming positive utilities and common default probabilities.

Theorem 2 *Let* $D : \mathcal{X} \times \mathcal{Y} \to \mathrm{I\!R}^+$ *denote a global distribution which is a weighted average of* k *local distributions* D_i*, all sharing the same default probabilities of classes. Considering a rule* $(A \to C) \in \mathcal{H}$*, let the marginal distributions of* D *and a local distribution* D_i $(i \in \{1, \ldots, k\})$ *be factor-similar up to* γ *for* A*, and let the joint distributions* D *and* D_i *be conditionally similar up to* ϵ *for the rule. Then the difference between global and local utilities of* $\mathrm{Q}^{(\alpha)}$ *is bounded by*

$$\begin{aligned}
&\max\left(0, \frac{\mathrm{Q}^{(\alpha)}_{D_i}(A \to C)}{\gamma^\alpha} - \frac{\epsilon}{\gamma^\alpha}\mathrm{Cov}_{D_i}(A \to C)^\alpha\right) \\
&\leq \max\left(0, \mathrm{Q}^{(\alpha)}_{D}(A \to C)\right) \\
&\leq \max\left(0, \gamma^\alpha \mathrm{Q}^{(\alpha)}_{D_i}(A \to C) + \epsilon\left[\gamma \mathrm{Cov}_{D_i}(A \to C)\right]^\alpha\right)
\end{aligned}$$

For valid choices of ϵ *these bounds are tight in general.*

For similarly distributed data the bounds are tight enough to allow for estimates with bounded uncertainty. A proof of theorem 2 and an illustration of the bounds are given in [8].

Please note that theorem 2 allows to exploit different estimates for each antecedent $A \subset \mathcal{X}$ under consideration. Hence, the theorem is not restricted to learning tasks in which conditional or marginal distributions are known to be very similar. It also allows to collect rule-specific bounds from various sites. Possible sources of rule-dependent bounds on γ and ϵ range from background knowledge over density estimates to previously cached queries.

The results given in this section also help to understand why methods like distributed boosting [6] are often successful in practice, although different kinds of skews at different sites are common, and are a known source of failure.

The question which rules do *not* allow to compute their utilities sufficiently well by techniques related to theorem 2 motivates a new extension of the learning task, discussed in Sec. 5, that explicitly takes the locality of data into account.

5. Relative local subgroup mining

As motivated in the last section, inhomogeneously distributed data allows to define subgroups as subsets of an example set[1] $\mathcal{E}_i$ that follow different distributions of the target attribute than $\mathcal{E}$. This definition of a subgroup has a natural interpretation that might be of practical interest in several domains. The corresponding rules could help to point

[1] More precisely, these definitions refer to the weight of subsets with respect to D and D_i. These weights are of course estimated based on the example sets.

out the characteristics of a single supermarket in contrast to the average supermarket, for example. The following utility function captures the idea of locally deviating rules.

Definition 8 *For $r \in \mathcal{H}$ the utility function $\text{RQ}_{D_i}^{(\alpha)}$ at a site i is defined as*

$$\text{RQ}_{D_i}^{(\alpha)}(r) := \text{COV}_{D_i}(r)^{\alpha} \cdot (\text{BIAS}_{D_i}(r) - \text{BIAS}_D(r))$$

The rules maximising this utility function are referred to as relative local subgroups.

Please note that only the global *conditional* distribution is required in this context, since COV is evaluated locally. Exploiting that COV differs by at most a factor of γ and assuming common class priors, theorem 2 can be simplified:

Corollary 1 *For a given target class C let*

$$\begin{aligned} rq_{max}^{(\alpha)} &:= \max\{\text{RQ}_{D_i}^{(\alpha)}(r) \mid r \in \mathcal{H},\ r \textit{ predicts } C\} \textit{ and} \\ rq_{min}^{(\alpha)} &:= \min\{\text{RQ}_{D_i}^{(\alpha)}(r) \mid r \in \mathcal{H},\ r \textit{ predicts } C\} \end{aligned}$$

denote the maximal and minimal utilities of relative local subgroups. Then for all rules $r' \in \mathcal{H}$ the difference between local and global utility is bounded by

$$\begin{aligned} & \gamma^{-\alpha} \cdot \left(\text{Q}_{D_i}^{(\alpha)}(r') - rq_{max}\right) \leq \text{Q}_D^{(\alpha)}(r') \\ & \leq \gamma^{\alpha} \cdot \left(\text{Q}_{D_i}^{(\alpha)}(r') - rq_{min}\right) \end{aligned}$$

if all terms are positive.

Cor. 1 allows to translate the utilities of local subgroups into global scores with bounded uncertainty for any rule-dependent γ. The special case of a common marginal distribution is obtained by setting $\gamma = 1$.

Corollary 2 *For $\gamma = 1$ the functions for local, relative local, and global subgroup discovery complete each other:*

$$\text{Q}_D^{(\alpha)}(A \to C) = \text{Q}_{D_i}^{(\alpha)}(A \to C) - \text{RQ}_{D_i}^{(\alpha)}(A \to C)$$

Obviously, the tasks of discovering relative local subgroups and that of approximating the global conditional distribution are of similar complexity in this case. Cor. 2 suggests how to detect global subgroups locally, given estimates of $\text{RQ}_{D_i}^{(\alpha)}$, and how to compute $\text{RQ}_{D_i}^{(\alpha)}$ from $\text{Q}_D^{(\alpha)}$ for $\gamma = 1$.

6. Conclusion

The behaviour of different rule selection metrics, their similarity for various skews and how well they may be estimated from samples has been investigated in recent years. What is lacking is an investigation of how these metrics behave in the scope of distributed learning. This paper is a first step into this direction. First of all it was shown that the utility measures common in the literature on subgroup discovery can be applied to homogeneously distributed data in the same way as to a single example set. If the different sites do not share a single underlying distribution generating the data, however, then even precise estimates may yield completely disjoint rulesets at all sites, none of which contains a single one of the best n rules. For the general case a tight bound for the difference between global and local rule utilities was proven, which allows to translate local rule utilities into global ones with bounded uncertainty. For the task of discovering rules that have a higher local than global utility it was shown that it is at least as hard as approximating the global conditional distribution of the target attribute. For a common marginal distribution one problem can be solved locally, given a solution for the other one.

The results indicate that distributed subgroup discovery is a hard problem, since it requires precise estimates of both, the global marginal and the global conditional distribution. Future work will evaluate algorithms for distributed rule mining, based on synthetic and real-world data.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft, Collaborative Research Centre 475 on *Reduction of Complexity for Multivariate Data Structures*.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large data bases. In *Proc. of the 20th Int. Conf. on Very Large Data Bases*, pages 478–499, 1994.

[2] D. Cheung and Y. Xiao. Effect of Data Skewness in Parallel Mining of Association Rules. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 48–60, 1998.

[3] P. A. Flach. The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics. In *Proc. of the 20th Int. Conf. on Machine Learning*. 2003.

[4] J. Fürnkranz and P. Flach. ROC 'n' Rule Learning – Towards a Better Understanding of Covering Algorithms. *Machine Learning*, 58(1):39–77, 2005.

[5] W. Klösgen. Explora: A Multipattern and Multistrategy Discovery Assistant. In *Advances in Knowledge Discovery and Data Mining*, chapter 3. AAAI Press/The MIT Press, 1996.

[6] A. Lazarevic and Z. Obradovic. Boosting algorithms for parallel and distributed learning. *Distributed and Parallel Databases Journal*, 11(2):203–229, 2002.

[7] T. Scheffer and S. Wrobel. Finding the Most Interesting Patterns in a Database Quickly by Using Sequential Sampling. *Journal of Machine Learning Research*, 3:833–862, 2002.

[8] M. Scholz. On the Complexity of Rule Discovery from Distributed Data. *SFB 475, Technical Report No. 31*, University of Dortmund, Germany, 2005.