Subgroup Discovery with Proper Scoring Rules

David Kaltenpoth

Abstract

The goal of subgroup discovery is to find interesting subgroups in our data. In the reviewed paper, Song et al. [12] combine proper scoring rules (PSR) with Bayesian generative modeling to obtain a theoretically well-founded score to evaluate the interestingness of subgroups which generalizes to future data. They show that the proposed approach outperforms a number of previous approaches on a variety of benchmark datasets and scores.

1 Subgroup Discovery

Subgroup Discovery (SGD) attempts to find interesting sub-populations in which the distribution of individuals on a set of attributes is distinct from that of the general population [3].

Since for smaller subgroups such statistical differences may arise due to pure chance, most approaches [3, 10] try to balance this bias by giving preference to larger subgroups.

However, many algorithms thereby prefer larger subgroups too strongly [12, 5], discarding other more distinct, but smaller, groups. In the worst case, this can even lead to inconsistent findings [5].

Song et al. [12] therefore propose to use proper scoring rules to appropriately balance the factors of distinctness and size, using a score of the form $m \cdot d(p, q)$, where m is the size of the subgroup and the divergence $d(p, q)$ measures the distinctiveness of the subgroup distribution from the global distribution.

Further, since the goal is to find subgroups which generalize to the future, Song et al. [12] further propose a generative approach to find a more robust estimate of interesting subgroups.

Lastly, they show that on both synthetic and real data, their methods tend to outperform previous ones on a number of different criteria, including the Weighted Relative Accuracy, the χ^2-statistic, as well as empirical information gain scores.

2 Measures of Interestingness

2.1 Measuring Subgroup Quality

Let our dataset $\{(X_i, Y_i) : i = 1, ..., n\}$ where the $X_i \in P$ describe attributes of our data and the Y_i the class to which point i belongs.

We describe subgroups by functions $\sigma : P \to \{0, 1\}$, where $\sigma(X_i) = 1$ means that the i-th data point belongs to subgroup σ with population Q. In the following, we will treat σ and Q interchangeably.

While in general the choice of the subgroup description language and the algorithmic issues with finding good or even optimal subgroups are hard in themselves (see e.g. [3, 9, 8]), here we do not concern ourselves with these problems.

Since our goal is to find interesting subgroups, we require a score $f(\sigma)$ to measure the goodness of such a subgroup. The task is then to find $\sigma^* = \text{argmax} f(\sigma)$.

If $Y_i = (Y_{i1}, ..., Y_{ik})$ encodes class $j \in \{1, ..., k\}$ as the vector $Y_i = e_j = (0, ..., 0, 1, 0, ..., 0)$ then we can write the global distribution for the data by p as $p = \overline{y} = \sum_{i=1}^{n} Y_i/n$.

The empirical distribution within subgroup σ is given by $q^\sigma = \overline{y}^\sigma = \sum_{i=1}^{n} \sigma(X_i) Y_i/m$, where $m = \sum_i \sigma(X_i)$ is the number of points belonging to σ.

One commonly used score is the weighted relative accuracy [10, 11]

$$f_{\text{WRAcc}}(\sigma) = m \sum_{j=1}^{k} \left| \overline{y}_j^\sigma - \overline{y}_j \right|, \tag{1}$$

which is possibly the simplest measure of distinctness between two distributions. However, as this does not have any nice interpretation in some theoretical framework of statistics, it is not clear why we should prefer this distance over any other one.

Given a subgroup σ and a measure q on σ we let $S^{\sigma,q,p}$ be the summary of the data which describes elements in σ by q and all other elements by p. That is, $S_i^{\sigma,q,p} = \sigma(X_i)q + (1 - \sigma(X_i))\, p$. We also write S^p if Q is empty (or the whole population).

2.2 Proper Scoring Rules

Given labels Y and a group σ, a *scoring rule* is any function $g(S, Y)$ assigning a value to the summary S given the label Y.

A *proper scoring rule* (PSR) $g(S, Y)$ is such that given labels Y and a group σ, the distribution q^σ is such that

$$q^\sigma = \text{argmin}_q g(S^{\sigma,q,p}, Y)$$

or equivalently

$$q^\sigma = \text{argmax}_q \left(g(S^p, Y) - g(S^{\sigma,q,p}, Y) \right)$$

in which case we can use $f(\sigma) := g(S^p, Y) - g(S^{\sigma,q^\sigma,p}, Y)$ as a score for g. We also write $g(p, Y)$ if our summary is given by a simple probability distribution, i.e. $g(p, Y) = g(S^p, Y)$.

What this means is that given a certain subgroup which we would like to describe, we can't do any better than by being honest about our beliefs of the underlying distribution. This makes PSRs as loss functions particularly popular with economists [13, 2].

We can therefore interpret $g(S, Y)$ as a loss function for a gamble where we use our summary/strategy S to make predictions.

To see why it isn't exactly trivial to find a proper scoring rule, assume that a ball is drawn from an urn, and you lose €1 for every single time you predict its color (red or blue) incorrectly. You quickly notice that the distribution of blue:red seems to be about $60 : 40$. But should you therefore say "blue" only 60% of the time? This wouldn't minimize your expected loss, as always saying "blue" gives an average loss of €0.40, while saying the truth gives an average loss of €0.48. See also Figure 1 for a comparison of the two strategies.

What, then, are loss functions which are proper?

The *logarithmic scoring rule* is given by

$$g_{LL}(S, Y) = -\sum_{i,j} Y_{ij} \log(S_{ij})$$

and the *Brier Score* is

$$g_B(S, Y) = \sum_{i,j} (Y_{ij} - S_{ij})^2 .$$

If $Y_1, ..., Y_n \sim \text{Cat}(p)$ are i.i.d. then in both cases, $p = \text{argmin}_p E_Y g(S^p, Y)$.

To come back to our example, for the logarithmic scoring rule the strategy of always saying "blue" corresponds to assigning a probability of 0 to "red" coming up. In particular, once "red" comes up you incur an infinite loss, and this is its expected value. A somewhat less extreme case is shown on the right of Figure 1.

One can show that for every PSR g there is a divergence d such that

$$E_{Y \sim p}(g(S^p, Y) - g(S^q, Y)) = d(p, q)$$

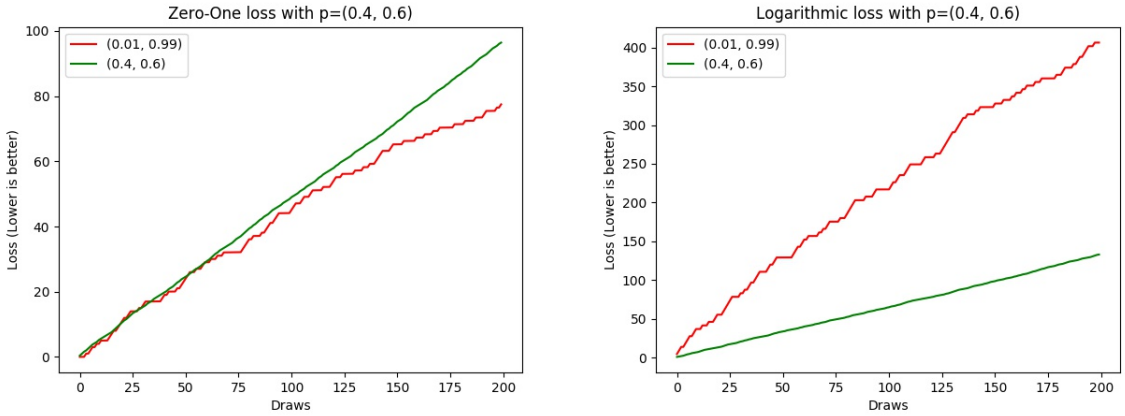

Figure 1: Comparing the strategies of nearly always picking the more probable choice with stating one's true belief. Left: Zero-One loss. Right: Logarithmic loss. As expected, stating one's true belief is the better choice for the proper scoring rule, but not otherwise.

For example, for g_{LL} respectively g_B the associated divergences are given by

$$d_{LL}(p,q) = \sum_{i=1}^{k} p_i \log(p_i/q_i)$$
$$d_B(p,q) = \sum_{i=1}^{k} (p_i - q_i)^2$$

the Kullback-Leibler (KL), respectively Euclidean, divergences.

One important feature of divergences in general is that $d(p,q) \geq 0$ with equality iff $p = q$. In general, d is not symmetric in its arguments.

For our urn example this means that saying anything other than "blue 60%, red 40%" leads to less than optimal outcomes, and the degree to which our choice is suboptimal is measured precisely by $d(p,q)$.

Particularly the KL divergence $d_{KL}(p,q)$ has a nice interpretation in information theory [6]: It is the amount of information we gain by using p instead of q. Put differently, it is the amount of information we lose if we try to approximate p using the distribution q.

For other divergences, including the Euclidean divergence above, a similar interpretation can generally not be given.

For more on the properties and importance of divergences, see [1].

3 Using Proper Scoring Rules to Find Subgroups

So how can we make use of PSRs to score the goodness of a subgroup?

We can use the *information gain* $g(S^p, Y) - g(S^{\sigma,q,p}, Y)$ due to changing the distribution on σ to a distribution q and keeping p outside σ.

We have seen above that since g is proper the distribution maximizing the information gain is $q^\sigma = \text{argmax}_q g(S^p, Y) - g(S^{\sigma,q,p}, Y)$. Since the only differences in predictions we make are for elements of the subgroup σ, and we are using empirical distributions, our score for σ is therefore

$$f_{\text{IG}}(\sigma) = g(S^p, Y) - g(S^{\sigma,q^\sigma,p}, Y) = m \cdot d(p, q^\sigma) \tag{2}$$

where d is the divergence associated with the the scoring rule.

As we have seen, the interpretation as information gain is particularly apt for the logarithmic scoring rule, due to its very direct relationship with entropy and the value of information [6].

Note that the form of (2) is of the same kind as that of (1) as well as many other scores [3]. The only difference is that here we use the divergence $d(p, q^\sigma)$ which is theoretically grounded instead of an ad hoc measure for distinctness of q^σ from p.

However, since σ can be a small subgroup (which is precisely what we want to allow), the distribution q^σ might, due to overfitting, not fit future data well. We therefore use a Bayesian generative approach to smooth q^σ. This will allow us to maximize the expected information gain for a new sample instead of the information gain on the data we have already got.

In particular, this contrasts with more classical methods which only look at the p-value of finding this score under the hypothesis that the group's distribution is no different from the global one.

3.1 Generative Models

To model the data-generating process we assume that each label comes from either the global distribution p, or, if the point belongs to the subgroup, from a distribution r.

We thus let $Z \sim \text{Bernoulli}(\gamma)$ describe whether a data point is chosen from the subgroup, if $Z = 1$, or from the global distribution, if $Z = 0$.

We further assume that the distribution on the subgroup σ is given by $r \sim \text{Dirichlet}(\beta)$ and the background distribution by p. Then

$$Y_1, ..., Y_m | Z, r \sim \text{Cat}(Zr + (1 - Z)p) \quad \text{i.i.d.}$$

The graphical representation for the model is shown in Figure 2

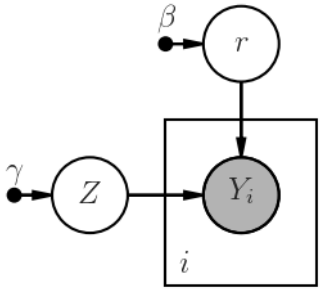

Figure 2: Graphical model proposed in Section 3.1.

Due to lack of further information we make our priors non-informative by setting $\gamma = 1/2, \beta = (1, ..., 1)$ [7]. Further let $C = \sum_{i=1}^{m} Y_i$ describe the number of times each class was observed.

Using the fact that the Dirichlet distribution is conjugate to the categorical distribution [4] one can show that the distribution maximizing the expected information gain conditional on $Z = 1, C = c$ is

$$\widehat{q} := \frac{c + \beta}{\sum_j (c_j + \beta_j)} = \text{argmax}_q E(g(p, Y) - g(q, Y) | Z = 1, C = c)\,.$$

This takes into account only our uncertainty about the distribution r generating the subgroup data.

Now if we further allow for uncertainty about whether the new data is drawn from the subgroup, i.e. we condition only on $C = c$ then

$$\widetilde{q} := a\frac{c + \beta}{\sum_j (c_j + \beta_j)} + (1 - a)p = \text{argmax}_q E(g(p, Y) - g(q, Y) | C = c)$$

where $a = P(Z = 1 | C = c)$. The expectations here are taken over Y with respect to the correct distributions as indicated by the conditionalization. The value of a can be computed directly from Bayes' theorem and our assumptions on the distributions of $Y_1, ..., Y_m | Z, r$ [4]. For detailed computations, see the appendix of Song et al. [12].

We see that both $\widehat{q}, \widetilde{q}$ are smoothed versions of q^σ taking into account our uncertainty about the generating process of the data.

We can now propose two new measures for the subgroup σ,

$$f_d(\sigma) := m \cdot d(\widehat{q}, p), \tag{3}$$

$$f_{PSR}(\sigma) := m \cdot d(\widetilde{q}, p) \tag{4}$$

where m is still the size of the subgroup.

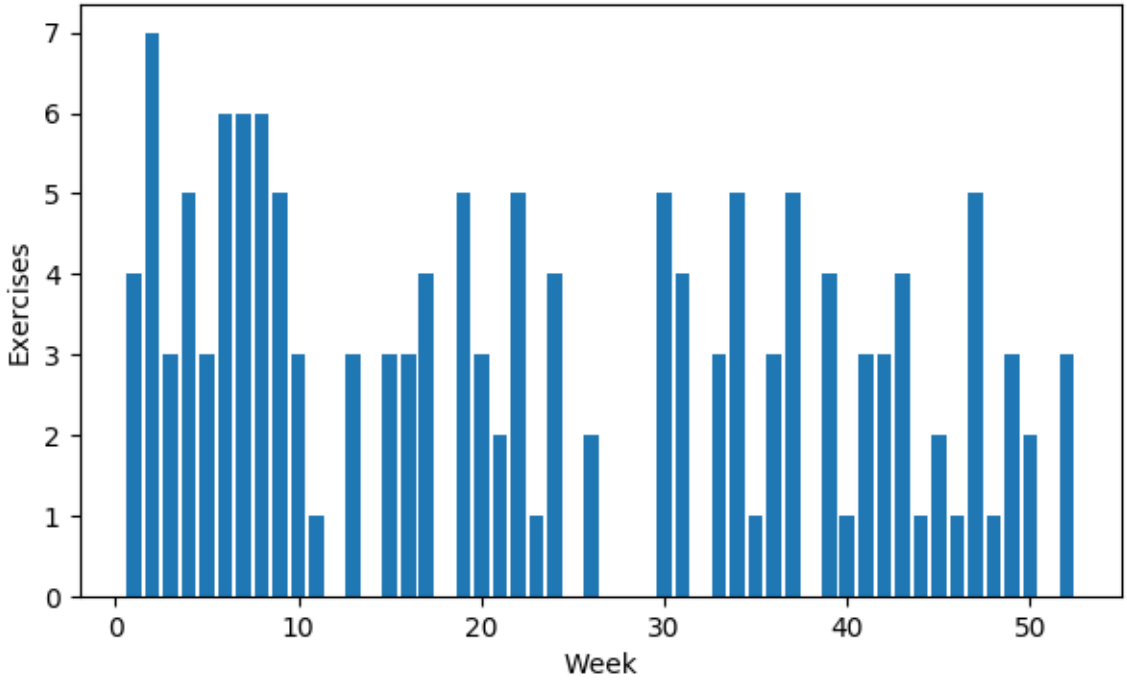

Figure 3: My exercise plans went really well. But only for the first few weeks.

4 Evaluation

If our wearable device tells us that we exercise properly 40% of the days, but that in the general population this is only 20% this could be due to at least two possible reasons.

While it is possible, that we are simply more fitness-conscious in general, it is also plausible that substantially more exercises were performed only for a certain time period of the year – like, say, the first month of the new year, just after we have decided to buy our new toy for Christmas – with no real differences for the rest of the year. See Figure 3 for an illustration.

Motivated by this, our feature space consists of the 52 weeks of the year, $\mathcal{X} = \{1, ..., 52\}$ and we use intervals of 2-8 weeks in our subgroup language. The subgroup is selected uniformly from all weeks, and our data generation process is as described in section 3.1.

It is convenient that in this setting it is possible to exhaustively check all possible subgroups to compare different methods, as testing with a heuristic algorithm might simply tell us which method works best with said heuristic algorithm.

The different scores we compare are the weighted relative accuracy (WRAcc) (1), IG-BS, IG-LL (obtained from (2)), as well as d-BS, d-LL, PSR-BS, PSR-LL (from (3),(4)).

Table 1: Averaged F-score on the artificial data, for different distributions of the classes. Best values are bold.

p_1	PSR-BS	PSR-LL	WRAcc	Chi2	IG-BS	IG-LL	d-BS	d-LL
.1	**.744**	.736	.597	.526	.030	.029	.742	.716
.2	.636	**.638**	.510	.436	.089	.091	.628	.631
.3	.587	**.589**	.480	.403	.218	.223	.581	.585
.4	.558	**.564**	.454	.390	.372	.379	.550	.559
.5	.567	**.569**	.458	.410	.561	.565	.561	.565

Table 2: Average logarithmic loss on the synthetic data for different distributions of the classes. Best values are bold.

p_1	PSR-BS	PSR-LL	WRAcc	Chi2	IG-BS	IG-LL	d-BS	d-LL
.1	**.344 ± .04**	**.344 ± .04**	.359 ± .04	.368 ± .04	.406 ± .06	.407 ± .06	**.344 ± .04**	.347 ± .04
.2	**.507 ± .03**	**.507 ± .03**	.517 ± .03	.520 ± .03	.539 ± .05	.540 ± .05	.508 ± .03	.509 ± .03
.3	**.610 ± .03**	**.610 ± .03**	.616 ± .02	.618 ± .02	.624 ± .03	.624 ± .03	.611 ± .03	.611 ± .03
.4	**.668 ± .02**	**.668 ± .02**	.673 ± .02	.674 ± .02	.671 ± .02	.671 ± .02	.670 ± .02	.669 ± .02
.5	.687 ± .02	**.686 ± .02**	.690 ± .01	.691 ± .01	.688 ± .02	.687 ± .02	.688 ± .02	.687 ± .02

We use F-score to compare different scores in their ability to find the correct subgroup. The results are shown in Table 1. We see that both PSR versions of our scores pretty consistently outperform all other methods with respect to F-score.

To test how well our scores do at predicting new data, we look at the loss of a new sample sampled from the same distribution. These scores are summarized in Table 2. We see that, at least for this artificial data set PSR versions are among the best regardless of the background distribution. Results for Brier score were almost identical.

5 Conclusion

In summary, Song et al. [12] propose to use PSRs, which theoretical foundations in information theory, to evaluate the distinctness of a subgroup.

To avoid finding subgroups which are very small, they then use a Bayesian generative approach which corresponds to smoothing the class distribution.

They show that not only do they in general find good subgroups as compared to the subgroup they do know exists, but their inferred summary also incurs a smaller predictive loss on future data sampled from the same distribution.

For the future, it would be nice to how well the approach works with numeric data.

Further, it would be interesting to see whether it is possible to find good optimization strategy for the subgroup score, e.g. using tight optimistic estimates [9].

References

[1] S. Amari. *Information geometry and its applications*. Vol. 194. Springer, 2016.

[2] O. Armantier and N. Treich. Eliciting beliefs: Proper scoring rules, incentives, stakes and hedging. *European Economic Review* **62** (2013), 17–40.

[3] M. Atzmueller. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **5** (2015), 35–49.

[4] C. M. Bishop. Pattern recognition. *Machine Learning* **128** (2006).

[5] M. Boley, B. R. Goldsmith, L. M. Ghiringhelli, and J. Vreeken. Identifying Consistent Statements about Numerical Data with Dispersion-Corrected Subgroup Discovery. *arXiv preprint arXiv:1701.07696* (2017).

[6] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2008.

[7] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Vol. 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

[8] H. Grosskreutz and D. Paurat. "Fast and Memory-Efficient Discovery of the Top-k Relevant Subgroups in a Reduced Candidate Space." In: Springer, Berlin, Heidelberg, 2011, 533–548.

[9] H. Grosskreutz, S. Rüping, and S. Wrobel. "Tight optimistic estimates for fast subgroup discovery." In: Springer, 2008, 440–456.

[10] N. Lavrač, P. Flach, and B. Zupan. "Rule evaluation measures: A unifying view." In: Springer, 1999, 174–185.

[11] N. Lavrač, B. Kavšek, P. Flach, and L. Todorovski. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research* **5** (2004), 153–188.

[12] H. Song, M. Kull, P. Flach, and G. Kalogridis. "Subgroup Discovery with Proper Scoring Rules." In: Springer, Cham, 2016, 492–510.

[13] G. Tziralis and I. Tatsiopoulos. Prediction markets: An extended literature review. *The journal of prediction markets* **1**.1 (2007), 75–91.