0.1 Optimizations

0.1.1 Stepwise Influence

In **??** we created a model of a system by taking the average value a feature has across groupings. In our example in **??** we were already able to see a problem **?**] described with group sampling. Features which share a group with an influential feature look more influential than they truly are. **?** describe a way to determine the influence more accuratly by determining the influence of a single parameter at a time.

The idea is to remove the influence an influential feature has before determining the influence of the next influenctial feature. We can do this in two steps, for each feature we want to determine the influence for. First, we need to find the most influential feature we have in out data. Then we have to estimate its influence and remove it from our data.

If we look at **??** we can use the average measurements of a feature to indentify the most influential. But, we have to be aware, that the average of the measurements still includes the baseline performance. By simply taking the highest measurement, we could end up picking the wrong feature, if, for example, the influential feature has a negative effect on the measurement. We want to pick the outlier value of the average feature measurements. One way to do this, is to pick the feature where the average measurement is the furthest away from the average of all measurements. In Table 1 we can see, the average of all average measurements would be approximatly 11.8, giving us F_6 as our outlier.

With the feature identified, we can remove it from our data. We estimate the influence of the feature the same way as previously, by taking the average of all its measurements. By substracting the value from all measurements where the feature was part of the group, we can get a better estimate of the influence the rest of the features have. We than remove the influential feature from our data and proceed with identifying the next influential feature. An example of this procedure can be found in Table 1. We identified F_6 as our most influential feature and removed it and its influences in other groups from the data. F_6 shared a group with F_5 and F_3 and after removing the estimated influence of F_6, both do not look as influenctial as in **??**.

In each round we identifiy the influence of one feature. With the influences of the features determined that way, we can build our model the same way as previously, but instead of using I_n in **??** and **??** we use the values determined by our stepwise analysis.

Table 1: Stepwise influence calculation

G_1	G_2	G_3	R	F_1	F_2	F_3	F_4	F_5	F_6	F_1	F_2	F_3	F_4	F_5	F_6
1	0	0	3	1	1	0	0	0	0	3	3				
0	1	0	6	0	0	1	1	0	0			6	6		
0	0	1	28	0	0	0	0	1	1					1,5	
1	0	0	2	1	0	0	1	0	0	2			2		
0	1	0	7	0	1	0	0	1	0		7			7	
0	0	1	25	0	0	1	0	0	1			-1,5			
										2,5	5	2,25	4	**4,25**	**26,5**

0.1.2 Feature coverage

In this work, we use a SAT-solver to create valid configurations. We can see the effects of it in **??**. Off-the-shelf SAT-solvers tend to find locally clustered solutions [**?**]. The repeating patterns in our samples are a result of it. The SAT-solver finds a solution to our constraints by changing as few variables as possible. Since at the beginning of each grouping, the constraints the solver has are similar, the SAT-solver gives a similar solution. This prevents us from making meaningful groupings, since features regularly share the same group.

We adapt the distance-based sampling strategy introduced by **?**] to help in creating more diverse groups. We use the Hamming distance as shown in **??** as our distance metric and the uniform distribution as our propability distribution. Instead of measruing the distance from the origin, we measure the distance from the first group of the previously created grouping.

In detail, at the start of each grouping we randomly pick a distance from the frist group created in the previous grouping. We than set the distance as a constraint for our SAT-solver to avoid getting a similar frist group. The following created groups are always dependent on the first group since they need to be distinct from each other. This way the groups created during each grouping do not follow a pattern of minimal change.

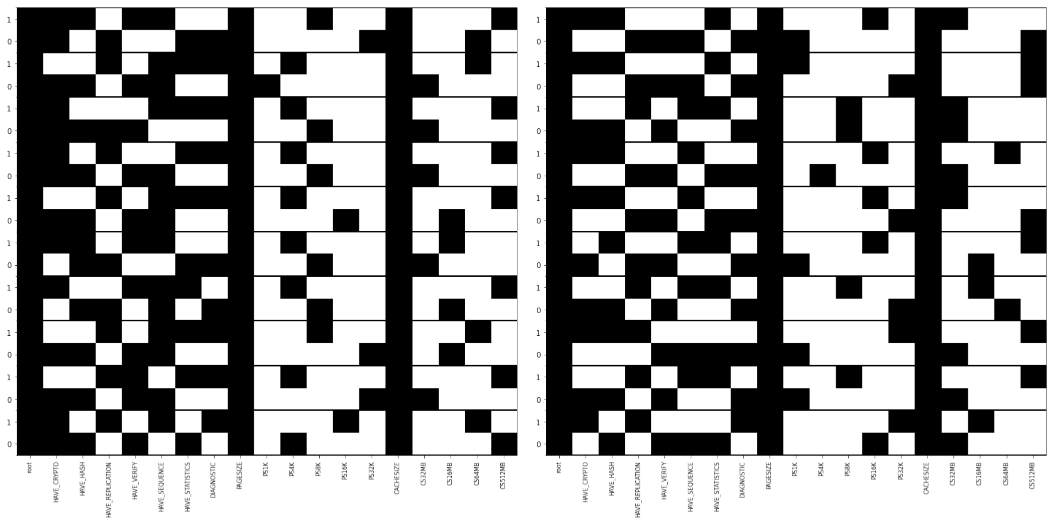

Figure 1: Samples generated on the BerkeleyDB Dataset by both variations with distance based group optimization. **Left:** Group Sampling - Hemming Distance - Group Size 2 - Groupings 10 **Right:** Group Sampling - Independent features - Group Size 2 - Groupings 10