What matters: Size does, Smarts Don't

Albrecht Zimmermann and Björn Bringmann

Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, 3001 Leuven, Belgium
{Albrecht.Zimmermann,Bjorn.Bringmann}@cs.kuleuven.be

Abstract. The usual data mining setting uses the full amount of data to derive patterns for different purposes. Taking cues from machine learning techniques, we explore ways to divide the data into subsets, mine patterns on them and use post-processing techniques for acquiring the result set. Using the patterns as features for a classification task to evaluate their quality, we compare the different subset compositions, and selection techniques. The two main results – that small independent sets are better suited than large amounts of data, and that uninformed selection techniques perform well – can to a certain degree be explained by quantitative characteristics of the derived pattern sets.

1 Introduction

When it comes to patterns to be used for classification, especially of complex data, class-correlating patterns are usually a good choice. By removing the influence of ad-hoc parameters such as frequency and confidence thresholds, found patterns can be expected not to fall prey to spurious phenomena in the data. No matter what the choice of patterns however, over-fitting effects are impossible to avoid and resulting patterns will of course influence the classifier. This becomes even more problematic if the amount of patterns is large, pairs and combinations of patterns reinforce each other's bias, or the classifier used to build the global model has limited safe-guards against over-fitting.

In a different work [2], we evaluated the effect of reducing redundancy between patterns on the accuracy of classifiers using those particular features. While we could show that reducing redundancy – in some cases rather strongly – did in fact improve accuracy, we adhered to what could be referred to as the standard *data mining* setting in that the underlying database was used as a whole for deriving the patterns we then filtered. Contrasting with this, the *machine learning* literature knows different techniques using parts of the labeled data for verification purposes of found patterns/built classifiers, such as re-sampling, validation sets, bootstrap sampling, and others.

We take a page out of the playbook of ML, first mining several sets of correlating patterns, and then using different criteria to create final result sets from these which are used as features for learning an SVM classifier. SVMs have inherent

capabilities for feature selection, and as a max-margin classifier more robustness regarding over-fitting effects, making them a good choice for evaluating feature quality.

The paper is structured as follows: In the next section, we explain the basic mechanisms for mining patterns, creating subsets of the data for mining and re-mining purposes, and lay out several selection methods for deriving the final result set. In Section 3, we report on the experimental evaluation of the proposed methods before concluding in Section 4.

2 Mining and Merging Correlating Patterns

The data we focus on in our evaluation is molecular data in the form of undirected graphs:

Definition 1 Graphs *An* undirected, labeled graph $\mathcal{G}(\mathbb{V}, \mathbb{E}, \lambda, \Sigma)$ *consists of a finite set* $\mathbb{V}$ *of vertices, a set* $\mathbb{E} \subseteq \{\{u, v\} | u, v \in \mathbb{V}, u \neq v\}$ *of edges, an alphabet* Σ *and a labeling function* $\lambda : (\mathbb{V} \cup \mathbb{E}) \rightarrow \Sigma$. *We denote the language of all possible graphs with* $\mathcal{L}_G$.

Our choice of pattern on graph-structured data are sequences, motivated by the observation that sequences are essentially as useful as features as are trees or graphs [3], while at the same time being far easier to mine.

Definition 2 Sequences *A* sequence *is a graph where no vertex has more than two edges (i.e. no branches), and has one more vertex than edges,* $\forall v \in \mathbb{V} : \left|\{\{v, u\} | u \in \mathbb{V}\}\right| \leq 2$ *and* $|\mathbb{E}| + 1 = |\mathbb{V}|$. *We denote the language of all possible sequences with* $\mathcal{L}_S$.

Given these languages, we define a predicate $match : \mathcal{L}_G \times \mathcal{L}_G \mapsto \{true, false\}$, and based on this define the notion of *generality*:

$$match(g, s) = \exists g' \text{ subgraph of } s \text{ with } g' \text{ isomorphic to } g$$

Definition 3 *Given two patterns* $p_g, p_s \in \mathcal{L}_G$, p_g *is said to be* more general than or equal to p_s, *denoted by* $p_g \preceq p_s$, *iff* $match(p_g, p_s)$. *If* $p_g \preceq p_s$ *holds but not* $p_s \preceq p_g$, p_g *is said to be* more general than p_s, *denoted* $p_g \prec p_s$.

Given a set of instances $\mathcal{D}_m = \{d_i | d_i \in \mathcal{L}_G\}$, each being labeled with one of the class labels $\{pos, neg\}$, we can use a correlation measure $\sigma(p, \mathcal{D}_m)$, such as χ^2, to mine the top-k patterns on $\mathcal{D}_m$, i.e. the k sequences p having the largest correlation with the class, according to σ [6]. Furthermore, to avoid redundancy in the resulting solution set, one typically employs patterns that are free:

Definition 4 Freeness *Given a data set* $\mathcal{D}_m$, *a pattern* p *is said to be* free *iff* $\nexists p', p' \prec p : \sigma(p', \mathcal{D}_m) = \sigma(p, \mathcal{D}_m)$.

The solutions to the mining task can be conveniently modeled using

$$\mathcal{T}h_k(\mathcal{D}_m) = \{p \mid free(p, \mathcal{D}_m), p \text{ among the } k\text{-best patterns w.r.t. } \sigma(p, \mathcal{D}_m)\}$$

As said before, this is the *standard* data mining setting, which we will use as a base-line technique. We compare the *quantitative* results (composition of the pattern sets) and *qualitative* results (predictive accuracies of the SVM) of the post-mining methods and the base-line. Different methods for selecting the final pattern set are described in the following sections.

2.1 Using a validation set

The most basic approach consist of using a certain fraction q of the total data $\mathcal{D}$ as the actual mining set $\mathcal{D}_m$, with size $q \cdot |\mathcal{D}|$. The rest would be used as a validation set $\mathcal{D}_{validate}$, of size $(1-q) \cdot |\mathcal{D}|$. This approach also means that there is need for *two* k-values, $k_m, k_{validate}$, with $k_m > k_{validate}$. After termination of the mining process on $\mathcal{D}_m$, the k_m patterns returned by the miner are evaluated on $\mathcal{D}_{validate}$ and re-ranked, according to the σ-score achieved on this validation set. From those the $k_{validate}$ best scoring patterns are returned to the user. This intuitively explains the second condition since for $k_m = k_{validate}$ the validation scores (and re-ranking) have no effect on the selection of patterns. An alternative would lie in only returning patterns that proved to be significant (according to some p-value) on the validation set.

This technique is related to the use of validation sets by approaches such as RIPPER [4] which use them to get a more realistic estimate of classification rules.

There are of course several interesting questions w.r.t. this approach. For instance regarding the proportion of training and validation data, and how to best determine it. Additionally, it would be worth investigating how the two k parameters relate, i.e. How much larger than $k_{validate}$ does k_m have to be for the composition of the top-$k_{validate}$ to stabilize? We plan to have a look into those topics in the further course of our work. Our focus in this paper is on a second technique that for now we find to be more intriguing.

2.2 Aggregating subset results

The second approach is somewhat connected to sub-sampling techniques such as *Bagging* [1] for alleviating over-fitting effects in classifier learning. In this technique, subsets $\mathcal{D}_m^i$ of $\mathcal{D}$ are created, and the top k_m patterns mined from them. The union of all such result sets $\Phi_{all} = \bigcup_i \mathcal{T}h_{k_m}(\mathcal{D}_m^i)$ has the property $|\Phi_{all}| \geq k_m$. All patterns $p \in \Phi_{all}$ are then re-evaluated according to some aggregation metric, and a subset (e.g. the top-k_m patterns) returned to the user.

There are two main decisions that influence the result of this approach, namely the choice of subsets and the aggregation metric used. The size of the final set to be returned is obviously also important but has less effect than the afore-mentioned two choices, we believe.

Composition of subsets The main difference between the two alternatives of forming subsets of $\mathcal{D}$ lies in whether there is overlap among subsets used or not. The straightest-forward approach consists of segmenting $\mathcal{D}$ into f folds $\mathcal{F}_i, \forall 0 \leq i, j \leq f, i \neq j : F_i \cap F_j = \emptyset$. We define $\hat{\mathcal{D}}^i_m = F_i$ and $\overline{\mathcal{D}}^i_m = \bigcup_{j \neq i} F_j$. In both cases *all* instances in the data have an effect on the final result with the same weight (in contrast to random subsampling).

In the first case, $\mathcal{D}^i_m = \hat{\mathcal{D}}^i_m$, the smaller amount of data used should lead to a faster mining operation, and the non-overlap would likely lead to a larger amount of patterns in $|\Phi_{all}|$, each more finely tuned to the particular fold. This might be an advantage in that patterns are mined that would be excluded when mining on larger amounts of data – but maybe their exclusion happened for good reason.

In the second case, $\mathcal{D}^i_m = \overline{\mathcal{D}}^i_m$, the effect that each instance is involved in several mining operations will likely lead to the mining of less, and more stable, patterns. Depending on the number of folds however, and therefore the size of $\overline{\mathcal{D}}^i_m$, the resulting set of patterns might not be much different from a mining operation on the full $\mathcal{D}$.

Aggregation metrics The goal of any aggregation metric used lies in ranking the patterns in Φ_{all} by using information from all subsets $\mathcal{D}^i_m$ mined on. To this end, the simplest metric would take the following form:

$$m_{count}(p) = |\{i \ : \ p \in \mathcal{T}h_{k_m}(\mathcal{D}^i_m)\}|$$

Basically m_{count} counts for each pattern p in how many of the $\mathcal{D}_m$ it was found among the top-k_m. The patterns are then ranked in decreasing order according to m_{count}. This measure does only check whether a pattern was mined at all, however, not what its particular rank was in the respective result sets. Hence, another metric, m_{rank} would consist of the following:

$$rank(p, \mathcal{D}) = \begin{cases} 1 + k_m - \inf\{k \mid p \in \mathcal{T}h_k(\mathcal{D})\} & \textit{if } \ p \in \mathcal{T}h_{k_m}(\mathcal{D}) \\ 0 \ \textit{otherwise} \end{cases}$$

$$m_{rank}(p) = \frac{1}{f} \sum_i rank(p, \mathcal{D}^i_m)$$

Again, all patterns in Φ_{all} would be re-ranked in decreasing order according to the metric.

A final metric, m_σ would take the form:

$$m_\sigma(p) = \sigma(p, \bigcup_i \mathcal{D}^i_m)$$

calculating the score for each pattern according to σ on the entire data set.

Selection criteria Due to the fact that $|\Phi_{all}| \geq k_m$, one can simply return the top-k_m after the re-ranking via one of the metrics $m(p)$. Thus, given a value k_m, a metric m, and a set of subsets $\mathcal{M} = \{\mathcal{D}^i_m\}$, our goal is to select

$$\varphi_{k_m}(\mathcal{M}, m) \subseteq \Phi_{all}$$

such that the $p_i \in \varphi_{k_m}(\mathcal{M}, m)$ are the k_m highest ranked pattern in Φ_{all} according to the measure m.

Additionally, this framework does allow for a second k-value (k'_m), similar to the one of the validation set approach which is used to define the size of the final result set. In this case $\Phi_{all} = \bigcup_i \mathcal{T}h_{k_m}(\mathcal{D}^i_m)$ but $\varphi_{k'_m}(\mathcal{M}, m) \subseteq \Phi_{all}$ is now such that the $p_i \in \varphi_{k'_m}(\mathcal{M}, m)$ are the k'_m highest ranked pattern in Φ_{all} according to the measure m.

Finally, the fact that the metrics assign new values to each pattern also means that other criteria can be used. One such criterion could be to use $m_{count}(p) \geq v$, meaning that a pattern has to be supported by at least v subsets to be selected. This is also a direction we intend to explore in future work.

3 Experimental Evaluation

For the experimental evaluation, we arbitrarily picked 8 data sets from the NCI-60 data set collection [7] (cf. Table 1), and mine sequential patterns on them. To balance the against the bias introduced by the partially overlapping NCI-60 sets, we additionally used a molecular data set where the set target variable is mutagenicity of compounds [5]. k_m was set to $10, 25, 50, 75, 100$, giving a reasonable range of values across which to compare. f took the values $3, 5, 7$, allowing for $\hat{\mathcal{D}}_m$ settings of different size and size difference to the $\overline{\mathcal{D}}_m$ settings. Since this is preliminary work, we set $k'_m = k_m$, and evaluated only k-best selection techniques, not threshold-based ones.

	instances			
Name	positive	negative	total	% pos
CCRF_CEM	2217	1263	3480	0,637%
COLO_205	1943	1702	3645	0,533%
786_0	1832	1674	3506	0,522%
CAKI_1	1865	1715	3580	0,520%
A498	1782	1698	3480	0,512%
A549_ATCC	1901	1833	3734	0,509%
ACHN	1795	1736	3531	0,508%
BT_549	1399	1379	2778	0,503%
Mutagenicity	2401	1936	4337	0,554%

Table 1. Properties of the 8 NCI Datasets and the Mutagenicity Dataset used, sorted by size of positive fraction

Two baseline techniques were also evaluated for comparison against: on the one hand the standard setting where the complete $\mathcal{D}$ is used for mining, which we identified as the standard data mining setting before. On the other hand we used a post-processing baseline where the selection method consists of picking k'_m patterns at random from Φ_{all} with uniform probability. Since this method does not use any information on the quality of individual patterns, nor their relationship with each other, we use it as a baseline to see whether the better informed methods enjoy an advantage.

To get a robust accuracy estimate, a 10-fold cross-validation was performed. All folds – both for accuracy and post-mining purposes – were stratified. As explained above, an SVM classifier was used for accuracy estimates. SVMs possess certain inherent feature selection capabilities, giving features that do not add relevant information small weights. In addition, an SVM attempts to find a separating hyperplane with a maximal margin to both classes, allowing it to find decision surfaces that should generalize better on unseen data than ones from classifiers that are lacking a regularization mechanism. Both of these characteristics make an SVM a good choice for evaluating the quality of a feature set. A potential argument against using an SVM is of course interpretability. However, the use of relatively large set of patterns in combination with the kind of techniques needed for other classifiers to achieve competitive accuracies to an SVM, e.g. *Boosting*, make interpreting such models difficult anyway. The SVM's C parameter was tuned via a 5-fold cross-validation on the training data, with potential values $2^i, i \in [-2, 14]$.

3.1 Quantitative results

In a first step, we report quantitative characteristics of Φ_{all} in Tables 2, 3, 4, and 5. Given that the results for the NCI-60 data sets are similar, we picked one at random and report on its characteristics. To this end we list $|\Phi_{all}|/k_m$ for the two alternatives regarding construction of the $\mathcal{D}_m$, the minimum and maximum m_{count} for patterns in Φ_{all}, the minimum and maximum m_{rank} for patterns in Φ_{all}, and $overlap = |\varphi_{k_m}(\mathcal{M}, m_\sigma) \cap \mathcal{T}h_{k_m}(\bigcup \mathcal{M})|$. What we would expect is the following:

- $|\Phi_{all,\hat{\mathcal{D}}_m}|/k_m \gg |\Phi_{all,\overline{\mathcal{D}}_m}|/k_m$ – Smaller $\mathcal{D}_m$ can be expected to lead to a larger variety of patterns
- $\min_{p \in \Phi_{all}} m_{count}(p) > 1$ – Any correlating patterns can be expected to appear in more than one result list
- $\max_{p \in \Phi_{all}} m_{count}(p) \approx f$ – The best correlating patterns can be expected to generalize over most $\mathcal{D}_m$
- $\min_{p \in \Phi_{all}} m_{rank}(p) > 1/f$ – Any correlating patterns can be expected to appear in more than one result list
- $\max_{p \in \Phi_{all}} m_{rank}(p) \approx k_m$ – The best correlating patterns can be expected to generalize over most $\mathcal{D}_m$, appearing with a high ranking
- $overlap_{\overline{\mathcal{D}}_m} \geq overlap_{\hat{\mathcal{D}}_m}$ – Larger $\mathcal{D}_m$ lead to results more similar to the standard setting

f	Overlap	$\lvert\Phi_{all}\rvert/k_m$	$\max_{p\in\Phi_{all}} m_{count}(p)$	$\max_{p\in\Phi_{all}} m_{rank}(p)$	$\min_{p\in\Phi_{all}} m_{rank}(p)$
			$k = 10$		
3	0,7 ± 0,949	2,000 ± 0,200	3 ± 0,000	7,800 ± 1,033	0,333 ± 0,000
5	0 ± 0,000	3,260 ± 0,302	4,2 ± 0,632	6,260 ± 0,766	0,200 ± 0,000
7	0 ± 0,000	4,630 ± 0,434	4,7 ± 0,483	4,814 ± 0,919	0,143 ± 0,000
			$k = 25$		
3	0,3 ± 0,483	2,136 ± 0,163	3 ± 0,000	22,800 ± 1,033	0,333 ± 0,000
5	0 ± 0,000	3,428 ± 0,204	4,8 ± 0,422	19,420 ± 1,459	0,200 ± 0,000
7	0 ± 0,000	4,668 ± 0,305	5,8 ± 0,632	15,614 ± 1,355	0,143 ± 0,000
			$k = 50$		
3	0,2 ± 0,632	2,114 ± 0,131	3 ± 0,000	47,800 ± 1,033	0,333 ± 0,000
5	0 ± 0,000	3,290 ± 0,208	5 ± 0,000	43,820 ± 2,165	0,200 ± 0,000
7	0 ± 0,000	4,454 ± 0,367	6,6 ± 0,699	36,700 ± 2,203	0,143 ± 0,000
			$k = 75$		
3	0,3 ± 0,949	2,056 ± 0,123	3 ± 0,000	72,800 ± 1,033	0,367 ± 0,105
5	0 ± 0,000	3,167 ± 0,171	5 ± 0,000	68,760 ± 2,299	0,200 ± 0,000
7	0 ± 0,000	4,389 ± 0,283	6,7 ± 0,675	60,171 ± 4,035	0,143 ± 0,000
			$k = 100$		
3	2,2 ± 4,662	2,018 ± 0,128	3 ± 0,000	97,800 ± 1,033	0,333 ± 0,000
5	0 ± 0,000	3,114 ± 0,171	5 ± 0,000	93,760 ± 2,299	0,200 ± 0,000
7	0 ± 0,000	4,321 ± 0,234	6,8 ± 0,422	84,071 ± 5,767	0,143 ± 0,000

Table 2. Quantitative characteristics for pattern sets mined on $\hat{\mathcal{D}}_m^i$ (NCI)

f	Overlap	$\lvert\Phi_{all}\rvert/k_m$	$\max_{p\in\Phi_{all}} m_{count}(p)$	$\max_{p\in\Phi_{all}} m_{rank}(p)$	$\min_{p\in\Phi_{all}} m_{rank}(p)$
			$k = 10$		
3	5,6 ± 1,506	1,440 ± 0,150	3 ± 0,000	9,033 ± 0,508	0,367 ± 0,105
5	5,7 ± 1,252	1,430 ± 0,125	5 ± 0,000	9,420 ± 0,416	0,260 ± 0,135
7	6 ± 1,764	1,400 ± 0,176	7 ± 0,000	9,500 ± 0,318	0,157 ± 0,045
			$k = 25$		
3	8,4 ± 3,098	1,644 ± 0,125	3 ± 0,000	24,033 ± 0,508	0,333 ± 0,000
5	8,5 ± 2,014	1,660 ± 0,080	5 ± 0,000	24,420 ± 0,416	0,220 ± 0,063
7	7,9 ± 2,885	1,684 ± 0,115	7 ± 0,000	24,500 ± 0,318	0,157 ± 0,045
			$k = 50$		
3	16,7 ± 4,715	1,648 ± 0,088	3 ± 0,000	49,033 ± 0,508	0,333 ± 0,000
5	18,4 ± 3,373	1,632 ± 0,067	5 ± 0,000	49,420 ± 0,416	0,200 ± 0,000
7	21,1 ± 3,315	1,578 ± 0,066	7 ± 0,000	49,500 ± 0,318	0,171 ± 0,090
			$k = 75$		
3	29,2 ± 6,356	1,599 ± 0,082	3 ± 0,000	74,033 ± 0,508	0,433 ± 0,161
5	34,1 ± 5,131	1,545 ± 0,068	5 ± 0,000	74,420 ± 0,416	0,280 ± 0,103
7	36,6 ± 5,522	1,512 ± 0,073	7 ± 0,000	74,500 ± 0,318	0,157 ± 0,045
			$k = 100$		
3	45,7 ± 7,484	1,535 ± 0,067	3 ± 0,000	99,033 ± 0,508	0,333 ± 0,000
5	52,1 ± 6,367	1,480 ± 0,064	5 ± 0,000	99,420 ± 0,416	0,320 ± 0,140
7	54,4 ± 8,208	1,456 ± 0,082	7 ± 0,000	99,500 ± 0,318	0,157 ± 0,045

Table 3. Quantitative characteristics for pattern sets mined on $\overline{\mathcal{D}}_m^i$ (NCI)

As the result tables show, most of our expectations hold for the NCI data, with the only serious exceptions being our assumptions about the "worst" patterns – which usually do not appear in more than one $\mathcal{T}h_k(\mathcal{D})$. This indicates that even when using correlation measures, different data sets quickly lead to differing (and incidentally less interpretable) mining results.

f	Overlap	$\lvert\Phi_{all}\rvert/k_m$	$\max_{p\in\Phi_{all}} m_{count}(p)$	$\max_{p\in\Phi_{all}} m_{rank}(p)$	$\min_{p\in\Phi_{all}} m_{rank}(p)$
			$k = 10$		
3	0,5 ± 0,972	2,34 ± 0,255	2,6 ± 0,516	6,467 ± 1,772	2,600 ± 0,211
5	0,3 ± 0,675	3,56 ± 0,414	3,6 ± 0,699	5,100 ± 0,682	1,880 ± 0,103
7	0,0 ± 0,000	4,92 ± 0,432	4,1 ± 0,994	4,329 ± 1,035	1,571 ± 0,151
			$k = 25$		
3	2,0 ± 1,700	2,124 ± 0,160	2,9 ± 0,316	19,767 ± 3,611	5,433 ± 0,738
5	1,0 ± 1,054	3,184 ± 0,318	4,4 ± 0,699	16,86 ± 2,409	4,580 ± 0,175
7	0,7 ± 1,252	3,824 ± 0,366	5,4 ± 0,516	13,171 ± 3,077	3,629 ± 0,518
			$k = 50$		
3	6,1 ± 4,677	2,118 ± 0,132	3,0 ± 0,000	43,933 ± 5,648	9,633 ± 0,597
5	2,9 ± 2,025	2,804 ± 0,202	4,9 ± 0,316	35,280 ± 6,153	7,400 ± 0,490
7	2,0 ± 2,582	2,824 ± 0,283	5,7 ± 0,675	21,671 ± 5,279	4,457 ± 0,908
			$k = 75$		
3	12,5 ± 6,835	2,07467 ± 0,115	3,0 ± 0,000	67,600 ± 5,760	13,800 ± 1,565
5	8,6 ± 5,758	2,212 ± 0,155	4,9 ± 0,3162	43,460 ± 6,872	7,380 ± 1,069
7	9,3 ± 6,651	1,956 ± 0,228	5,7 ± 0,675	22,586 ± 5,522	3,129 ± 0,977
			$k = 100$		
3	16,7 ± 8,883	1,963 ± 0,161	3,0 ± 0,000	87,400 ± 9,597	16,767 ± 3,236
5	18,5 ± 6,721	1,707 ± 0,121	4,9 ± 0,316	44,640 ± 7,168	5,260 ± 0,938
7	22,5 ± 8,708	1,467 ± 0,171	5,7 ± 0,675	22,586 ± 5,522	1,900 ± 0,788

Table 4. Quantitative characteristics for pattern sets mined on $\hat{\mathcal{D}}_m^i$ (Mutagenicity)

On the mutagenicity data, the results look somewhat different however. On the one hand are the overlap results not as extreme – there is quite a bit overlap even for the $\hat{\mathcal{D}}_m$ settings, while the overlap for $\overline{\mathcal{D}}_m$ settings is smaller than on the NCI data. On the other hand do the subsets have a stronger effect on the ranking of patterns – "best" patterns are less often in all top-k than for the NCI data, and do not always come out ranked highest, while there are patterns that are only in one result list but relatively highly ranked, and "worst" patterns that can be found in several result lists. Generally, it can be observed that splitting the data into subsets leads to larger churn among patterns found.

3.2 Qualitative results

After having seen the quantitative characteristics of the different pattern sets the probably more interesting question is which of the proposed mining- and post-mining techniques select patterns which are useful in representing data

f	Overlap	$\lvert\Phi_{all}\rvert/k_m$	$\max_{p\in\Phi_{all}} m_{count}(p)$	$\max_{p\in\Phi_{all}} m_{rank}(p)$	$\min_{p\in\Phi_{all}} m_{rank}(p)$
			$k = 10$		
3	1,3 ± 1,567	1,710 ± 0,338	2,6 ± 0,516	6,467 ± 1,772	2,600 ± 0,211
5	1,3 ± 1,703	1,750 ± 0,375	3,6 ± 0,699	5,100 ± 0,682	1,880 ± 0,103
7	1,3 ± 1,567	1,740 ± 0,386	4,1 ± 0,994	4,329 ± 1,035	1,571 ± 0,151
			$k = 25$		
3	3,7 ± 2,110	1,464 ± 0,128	2,9 ± 0,316	19,767 ± 3,611	5,433 ± 0,738
5	3,6 ± 1,955	1,416 ± 0,112	4,4 ± 0,699	16,860 ± 2,409	4,580 ± 0,175
7	4,0 ± 2,000	1,364 ± 0,111	5,4 ± 0,516	13,171 ± 3,077	3,629 ± 0,518
			$k = 50$		
3	11,3 ± 3,466	1,542 ± 0,093	3,0 ± 0,000	43,933 ± 5,648	9,633 ± 0,597
5	11,7 ± 3,561	1,506 ± 0,106	4,9 ± 0,316	35,280 ± 6,153	7,400 ± 0,490
7	12,9 ± 3,755	1,494 ± 0,120	5,7 ± 0,675	21,671 ± 5,279	4,457 ± 0,908
			$k = 75$		
3	22,6 ± 3,950	1,564 ± 0,095	3,0 ± 0,000	67,600 ± 5,760	13,80 ± 1,565
5	24,3 ± 4,111	1,508 ± 0,126	4,9 ± 0,316	43,460 ± 6,872	7,380 ± 1,069
7	25,8 ± 4,492	1,503 ± 0,096	5,7 ± 0,675	22,586 ± 5,522	3,129 ± 0,977
			$k = 100$		
3	34,8 ± 7,554	1,569 ± 0,090	3,0 ± 0,000	87,400 ± 9,597	16,767 ± 3,236
5	38,7 ± 6,378	1,516 ± 0,132	4,9 ± 0,316	44,640 ± 7,168	5,260 ± 0,938
7	39,1 ± 7,593	1,514 ± 0,141	5,7 ± 0,675	22,586 ± 5,522	1,900 ± 0,788

Table 5. Quantitative characteristics for pattern sets mined on $\overline{\mathcal{D}}_m^i$ (Mutagenicity)

for classification purposes. As mentioned above, we used an SVM and 10-fold cross-validation to estimate the quality of pattern sets. Due to the fact that differences in accuracy were almost never significant, we omit the actual accuracy estimates here. Instead we report how the different methods (each time a combination of $\mathcal{D}_m$ composition and selection technique) compare giving a fixed $k_m = 10, 25, 50, 75, 100$ and $f = 3, 5, 7$ (Tables 6,7,8, and 9, 10, 11, respectively). Note that the tables for the NCI data show the aggregated wins for each approach, with more detailed tables reporting pair-wise comparisons in the appendix.

Each number denotes how often a particular technique has performed better than any other technique on any data set. Since 8 NCI data sets were used, giving a technique maximally 8 wins against any single other technique, and 9 techniques were evaluated, any given approach can have a maximum of 64 wins in Tables 6, 7, 8. Bold values denote the best-performing technique for a given k_m, while a circle ($\circ$) shows for which amount of patterns a given technique performed best.

The first, somewhat surprising insight, is that using large, overlapping $\mathcal{D}_m$ does not lead to good pattern selection. $\overline{\mathcal{D}}_m$ settings never perform best for a given k_m and usually perform better if only relatively few patterns are selected, suggesting that resampling does too little to counteract bias. Given that re-

	$k=10$	$k=25$	$k=50$	$k=75$	$k=100$
$\hat{\mathcal{D}}_m$, chi	10	7	10	13	18 ∘
$\hat{\mathcal{D}}_m$, random	38	39	44	43	52 ∘
$\hat{\mathcal{D}}_m$, rank	**49** ∘	42	39	**48**	31
$\hat{\mathcal{D}}_m$, top	42	**52** ∘	**49**	38	33
$\overline{\mathcal{D}}_m$, chi	18 ∘	18 ∘	10	8	14
$\overline{\mathcal{D}}_m$, random	25	48 ∘	44	41	32
$\overline{\mathcal{D}}_m$, rank	36	23	42 ∘	32	29
$\overline{\mathcal{D}}_m$, top	41 ∘	39	25	38	25
baseline	29	20	25	27	**54** ∘

Table 6. Accuracy wins for different k_m-values ($f = 3$)

	$k=10$	$k=25$	$k=50$	$k=75$	$k=100$
$\hat{\mathcal{D}}_m$, chi	9	2	14	16	22 ∘
$\hat{\mathcal{D}}_m$, random	**52**	**54**	50	**59** ∘	47
$\hat{\mathcal{D}}_m$, rank	**52**	49	**52**	51	**59** ∘
$\hat{\mathcal{D}}_m$, top	51 ∘	42	46	43	36
$\overline{\mathcal{D}}_m$, chi	14	28 ∘	19	8	4
$\overline{\mathcal{D}}_m$, random	32	43 ∘	34	36	28
$\overline{\mathcal{D}}_m$, rank	28	17	26	32	34 ∘
$\overline{\mathcal{D}}_m$, top	22	29 ∘	28	22	21
baseline	28	24	19	21	37 ∘

Table 7. Accuracy wins for different k_m-values ($f = 5$)

	$k=10$	$k=25$	$k=50$	$k=75$	$k=100$
$\hat{\mathcal{D}}_m$, chi	8	6	15	24	29 ∘
$\hat{\mathcal{D}}_m$, random	43	49	**54**	**56** ∘	**50**
$\hat{\mathcal{D}}_m$, rank	42	**54** ∘	52	50	**50**
$\hat{\mathcal{D}}_m$, top	**46** ∘	41	37	37	32
$\overline{\mathcal{D}}_m$, chi	22 ∘	19	11	8	5
$\overline{\mathcal{D}}_m$, random	19	46 ∘	43	34	37
$\overline{\mathcal{D}}_m$, rank	38 ∘	29	33	26	32
$\overline{\mathcal{D}}_m$, top	32 ∘	25	22	28	18
baseline	38 ∘	19	21	25	35

Table 8. Accuracy wins for different k_m-values ($f = 7$)

sampling forms the basis for, e.g., BAGGING techniques, we did not expect this outcome.

What can also be noticed is that the standard approach – using all training data for mining – produces suboptimal pattern sets. Only once this baseline approach is best, for $f = 3$, meaning relatively large folds where the *informed* selection techniques such as *count* and *rank* do not enjoy a large advantage. And even there it is closely followed by the random pattern selection - essentially the most uninformed one. Which in turn means that an unwritten paradigm of data mining – that using large amounts of data to the fullest will produce meaningful patterns – turns out to be questionable in this case.

The random technique is in fact the big winner of the entire comparison on the NCI data, given its naïve selection method. Although the reduction of redundancy using this technique is entirely by chance, it performs well in 4 of 15 settings, tying *rank* twice, which wins 5 times, with *count* after it at 3 wins. So the information which patterns generalize well over different subsets does not give a strong advantage in our case study.

In the case of the mutagenicity data set, the maximum number of wins possible is of course eight, given that nine techniques are compared pairwise on a single data set. There is for each value-combination of f and k_m always one approach that reaches this maximum, and again it's the random, *rank* and *count* techniques on the $\hat{\mathcal{D}}_m$ setting that share the honor, with the exception being the *rank* approach for $k_m = 10$ that performs best in the $\overline{\mathcal{D}}_m$ setting. Again, there are 15 combinations of f, and k_m-value and the *rank* approach does best in six of these, followed by the random technique with five, and *count* with four settings. These results are consistent with the results seen on the NCI data, suggesting that this is be a phenomenon that might not be restricted to a particular data set.

4 Conclusions

In this work, we investigated ways of using data sets for data mining with the goal of producing good features for classification of complex data. Two main insights arise from the experimental evaluation: 1) size **does** matter! Just not in the way usually assumed. Neither the standard data mining setting which uses large data sets to smooth over-fitting effects by pure size, nor re-sampling techniques that produce overlapping mining sets with the goal of uncovering the true underlying phenomena proved to be the most effective use of data. The best way we observed to use the data consisted of splitting it up into small, *independent* data subsets instead. 2) the actual selection matters far less than could be expected. Given a large enough variety of patterns, picking patterns at random proved to be a rather effective technique, as proved the average rank selector, which in effect picks patterns that were highly ranked at least once, even if in other subsets they were not.

As we have outlined in preceding sections already, this is preliminary work and there are several other techniques which we plan to evaluate. Potentially

	$k=10$	$k=25$	$k=50$	$k=75$	$k=100$
$\hat{\mathcal{D}}_m$, chi	1	1	1	1	1
$\hat{\mathcal{D}}_m$, random	5	**8**∘	**8**∘	**8**∘	5
$\hat{\mathcal{D}}_m$, rank	3	6	5	5	**8**∘
$\hat{\mathcal{D}}_m$, top	6	5	4	7∘	7∘
$\overline{\mathcal{D}}_m$, chi	2	2	2	2	2
$\overline{\mathcal{D}}_m$, random	4	4	7∘	6	6
$\overline{\mathcal{D}}_m$, rank	**8**∘	7	6	4	4
$\overline{\mathcal{D}}_m$, top	7∘	3	3	3	3
baseline	0	0	0	0	0

Table 9. Accuracy wins for different k_m-values ($f = 3$)

	$k=10$	$k=25$	$k=50$	$k=75$	$k=100$
$\hat{\mathcal{D}}_m$, chi	1	1	1	1	1
$\hat{\mathcal{D}}_m$, random	3	3	5	5	**8**∘
$\hat{\mathcal{D}}_m$, rank	5	5	7	**8**∘	7
$\hat{\mathcal{D}}_m$, top	**8**∘	**8**∘	**8**∘	7	5
$\overline{\mathcal{D}}_m$, chi	2	2	2	2	2
$\overline{\mathcal{D}}_m$, random	6∘	4	3	6∘	6∘
$\overline{\mathcal{D}}_m$, rank	7∘	7∘	6	4	3
$\overline{\mathcal{D}}_m$, top	4	6∘	4	3	4
baseline	0	0	0	0	0

Table 10. Accuracy wins for different k_m-values ($f = 5$)

	$k=10$	$k=25$	$k=50$	$k=75$	$k=100$
$\hat{\mathcal{D}}_m$, chi	1	1	1	1	1
$\hat{\mathcal{D}}_m$, random	2	3	6	7	**8**∘
$\hat{\mathcal{D}}_m$, rank	**8**∘	**8**∘	**8**∘	6	7
$\hat{\mathcal{D}}_m$, top	7	7	7	**8**∘	5
$\overline{\mathcal{D}}_m$, chi	3	2	2	2	2
$\overline{\mathcal{D}}_m$, random	5	5	5	5	6∘
$\overline{\mathcal{D}}_m$, rank	6∘	6∘	3	4	3
$\overline{\mathcal{D}}_m$, top	4	4	4	3	4
baseline	0	0	0	0	0

Table 11. Accuracy wins for different k_m-values ($f = 7$)

most significant is the fact that our preliminary results show that the one defining characteristic that qualifies a pattern set as a good feature set for classification is non-redundancy. Given the effectiveness of using independent subsets, a viable strategy should be to split up the initial data set into subsets according to pattern-effects, and re-iterate the mining process – effectively using the same effect that decision trees employ.

References

1. Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.
2. Björn Bringmann and Albrecht Zimmermann. The chosen few: On identifying valuable patterns. In *ICDM*, pages 63–72. IEEE Computer Society, 2007.
3. Björn Bringmann, Albrecht Zimmermann, Luc De Raedt, and Siegfried Nijssen. Don't be afraid of simpler patterns. In Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou, editors, *PKDD*, volume 4213 of *Lecture Notes in Computer Science*, pages 55–66. Springer, 2006.
4. William W. Cohen. Fast effective rule induction. In Armand Prieditis and Stuart J. Russell, editors, *Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123, Tahoe City, California, USA, July 1995. Morgan Kaufmann.
5. J. Kazius, S. Nijssen, J.N. Kok, T. Back, and A.P. IJzerman. Substructure mining using elaborate chemical representation. *Journal of Chemical Information and Modeling*, 46(2):597–605, 2006.
6. Shinichi Morishita and Jun Sese. Traversing itemset lattices with statistical metric pruning. In *PODS*, pages 226–236, Dallas, Texas, USA, May 2000. ACM.
7. Sanjay Joshua Swamidass, Jonathan H. Chen, Jocelyne Bruand, Peter Phung, Liva Ralaivola, and Pierre Baldi. Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. In *ISMB (Supplement of Bioinformatics)*, pages 359–368, 2005.

A Detailed Accuracy Wins for Different k_m-values

k=10, sf=3	baseline	chi, $\hat{\mathcal{D}}_m$	chi, $\overline{\mathcal{D}}_m$	random, $\hat{\mathcal{D}}_m$	random, $\overline{\mathcal{D}}_m$	rank, $\hat{\mathcal{D}}_m$	rank, $\overline{\mathcal{D}}_m$	top, $\hat{\mathcal{D}}_m$	top, $\overline{\mathcal{D}}_m$	Σ
baseline		**7**	**6**	3	**5**	2	2	3	1	**29**
chi, $\hat{\mathcal{D}}_m$	1		3	1	1	1	1	1	1	**10**
chi, $\overline{\mathcal{D}}_m$	2	**5**		2	3	1	2	1	2	**18**
random, $\hat{\mathcal{D}}_m$	**5**	**7**	**6**		**6**	3	**5**	3	3	**38**
random, $\overline{\mathcal{D}}_m$	3	**7**	**5**	2		1	3	1	3	**25**
rank, $\hat{\mathcal{D}}_m$	**6**	**7**	**7**	**5**	**7**		**5**	**6**	**6**	**49**
rank, $\overline{\mathcal{D}}_m$	**6**	**7**	**6**	3	**5**	3		3	3	**36**
top, $\hat{\mathcal{D}}_m$	**5**	**7**	**7**	**5**	**7**	2	**5**		4	**42**
top, $\overline{\mathcal{D}}_m$	**7**	**7**	**6**	**5**	**5**	2	**5**	4		**41**

Table 12. Wins for $k = 10$ on 3 subfolds

k=10, sf=5	baseline	chi, $\hat{\mathcal{D}}_m$	chi, $\overline{\mathcal{D}}_m$	random, $\hat{\mathcal{D}}_m$	random, $\overline{\mathcal{D}}_m$	rank, $\hat{\mathcal{D}}_m$	rank, $\overline{\mathcal{D}}_m$	top, $\hat{\mathcal{D}}_m$	top, $\overline{\mathcal{D}}_m$	Σ
baseline		**7**	**6**	1	4	0	4	0	**6**	**28**
chi, $\hat{\mathcal{D}}_m$	1		3	1	1	1	1	0	1	**9**
chi, $\overline{\mathcal{D}}_m$	2	**5**		0	1	1	2	1	2	**14**
random, $\hat{\mathcal{D}}_m$	**7**	**7**	**8**		**7**	4	**7**	**5**	**7**	**52**
random, $\overline{\mathcal{D}}_m$	4	**7**	**7**	1		2	4	2	**5**	**32**
rank, $\hat{\mathcal{D}}_m$	**8**	**7**	**7**	4	**6**		**8**	4	**8**	**52**
rank, $\overline{\mathcal{D}}_m$	4	**7**	**6**	1	4	0		1	**5**	**28**
top, $\hat{\mathcal{D}}_m$	**8**	**8**	**7**	3	**6**	4	**7**		**8**	**51**
top, $\overline{\mathcal{D}}_m$	2	**7**	**6**	1	3	0	3	0		**22**

Table 13. Wins for $k = 10$ on 5 subfolds

k=10, sf=7	baseline	chi, $\hat{\mathcal{D}}_m$	chi, $\overline{\mathcal{D}}_m$	random, $\hat{\mathcal{D}}_m$	random, $\overline{\mathcal{D}}_m$	rank, $\hat{\mathcal{D}}_m$	rank, $\overline{\mathcal{D}}_m$	top, $\hat{\mathcal{D}}_m$	top, $\overline{\mathcal{D}}_m$	Σ
baseline		**7**	**6**	3	**6**	4	4	2	**6**	**38**
chi, $\hat{\mathcal{D}}_m$	1		2	0	1	1	1	1	1	**8**
chi, $\overline{\mathcal{D}}_m$	2	**6**		1	**5**	1	2	2	3	**22**
random, $\hat{\mathcal{D}}_m$	**5**	**8**	**7**		**7**	4	4	4	4	**43**
random, $\overline{\mathcal{D}}_m$	2	**7**	3	1		1	2	1	2	**19**
rank, $\hat{\mathcal{D}}_m$	4	**7**	**7**	4	**7**		4	4	**5**	**42**
rank, $\overline{\mathcal{D}}_m$	4	**7**	**6**	4	**6**	4		2	**5**	**38**
top, $\hat{\mathcal{D}}_m$	**6**	**7**	**6**	4	**7**	4	**6**		**6**	**46**
top, $\overline{\mathcal{D}}_m$	2	**7**	**5**	4	**6**	3	3	2		**32**

Table 14. Wins for $k = 10$ on 7 subfolds

k=25, sf=3	baseline	chi, $\hat{\mathcal{D}}_m$	chi, $\overline{\mathcal{D}}_m$	random, $\hat{\mathcal{D}}_m$	random, $\overline{\mathcal{D}}_m$	rank, $\hat{\mathcal{D}}_m$	rank, $\overline{\mathcal{D}}_m$	top, $\hat{\mathcal{D}}_m$	top, $\overline{\mathcal{D}}_m$	Σ
baseline		**6**	4	1	2	1	4	0	2	**20**
chi, $\hat{\mathcal{D}}_m$	2		2	0	1	0	1	0	1	**7**
chi, $\overline{\mathcal{D}}_m$	4	**6**		2	1	1	2	0	2	**18**
random, $\hat{\mathcal{D}}_m$	**7**	**8**	**6**		2	4	**5**	4	3	**39**
random, $\overline{\mathcal{D}}_m$	**6**	**7**	**7**	**6**		**5**	**7**	4	**6**	**48**
rank, $\hat{\mathcal{D}}_m$	**7**	**8**	**7**	4	3		**7**	2	4	**42**
rank, $\overline{\mathcal{D}}_m$	4	**7**	**6**	3	1	1		0	1	**23**
top, $\hat{\mathcal{D}}_m$	**8**	**8**	**8**	4	4	**6**	**8**		**6**	**52**
top, $\overline{\mathcal{D}}_m$	**6**	**7**	**6**	**5**	2	4	**7**	2		**39**

Table 15. Wins for $k = 25$ on 3 subfolds

k=25, sf=5	baseline	chi, $\hat{\mathcal{D}}_m$	chi, $\overline{\mathcal{D}}_m$	random, $\hat{\mathcal{D}}_m$	random, $\overline{\mathcal{D}}_m$	rank, $\hat{\mathcal{D}}_m$	rank, $\overline{\mathcal{D}}_m$	top, $\hat{\mathcal{D}}_m$	top, $\overline{\mathcal{D}}_m$	Σ
baseline		**8**	4	1	1	1	4	2	3	**24**
chi, $\hat{\mathcal{D}}_m$	0		0	1	0	0	1	0	0	**2**
chi, $\overline{\mathcal{D}}_m$	4	**8**		1	1	2	**6**	2	4	**28**
random, $\hat{\mathcal{D}}_m$	**7**	**7**	**7**		**6**	**5**	**8**	**7**	**7**	**54**
random, $\overline{\mathcal{D}}_m$	**7**	**8**	**7**	2		3	**7**	4	**5**	**43**
rank, $\hat{\mathcal{D}}_m$	**7**	**8**	**6**	3	**5**		**8**	**5**	**7**	**49**
rank, $\overline{\mathcal{D}}_m$	4	**7**	2	0	1	0		1	2	**17**
top, $\hat{\mathcal{D}}_m$	**6**	**8**	**6**	1	4	3	**7**		**7**	**42**
top, $\overline{\mathcal{D}}_m$	**5**	**8**	4	1	3	1	**6**	1		**29**

Table 16. Wins for $k = 25$ on 5 subfolds

k=25, sf=7	baseline	chi, $\hat{\mathcal{D}}_m$	chi, $\overline{\mathcal{D}}_m$	random, $\hat{\mathcal{D}}_m$	random, $\overline{\mathcal{D}}_m$	rank, $\hat{\mathcal{D}}_m$	rank, $\overline{\mathcal{D}}_m$	top, $\hat{\mathcal{D}}_m$	top, $\overline{\mathcal{D}}_m$	Σ
baseline		**7**	**5**	1	1	0	2	1	2	**19**
chi, $\hat{\mathcal{D}}_m$	1		1	0	1	0	1	1	1	**6**
chi, $\overline{\mathcal{D}}_m$	3	**7**		1	1	0	2	2	3	**19**
random, $\hat{\mathcal{D}}_m$	**7**	**8**	**7**		4	**5**	**6**	**5**	**7**	**49**
random, $\overline{\mathcal{D}}_m$	**7**	**7**	**7**	4		3	**6**	**6**	**6**	**46**
rank, $\hat{\mathcal{D}}_m$	**8**	**8**	**8**	3	**5**		**8**	**6**	**8**	**54**
rank, $\overline{\mathcal{D}}_m$	**6**	**7**	**6**	2	2	0		1	**5**	**29**
top, $\hat{\mathcal{D}}_m$	**7**	**7**	**6**	3	2	2	**7**		**7**	**41**
top, $\overline{\mathcal{D}}_m$	**6**	**7**	**5**	1	2	0	3	1		**25**

Table 17. Wins for $k = 25$ on 7 subfolds

k=50, sf=3	baseline	chi, $\hat{\mathcal{D}}_m$	chi, $\overline{\mathcal{D}}_m$	random, $\hat{\mathcal{D}}_m$	random, $\overline{\mathcal{D}}_m$	rank, $\hat{\mathcal{D}}_m$	rank, $\overline{\mathcal{D}}_m$	top, $\hat{\mathcal{D}}_m$	top, $\overline{\mathcal{D}}_m$	Σ
baseline		**7**	**6**	2	2	1	1	2	4	**25**
chi, $\hat{\mathcal{D}}_m$	1		4	0	1	1	1	1	1	**10**
chi, $\overline{\mathcal{D}}_m$	2	4		1	0	0	0	1	2	**10**
random, $\hat{\mathcal{D}}_m$	**6**	**8**	**7**		4	**5**	4	3	**7**	**44**
random, $\overline{\mathcal{D}}_m$	**6**	**7**	**8**	4		**5**	**5**	3	**6**	**44**
rank, $\hat{\mathcal{D}}_m$	**7**	**7**	**8**	3	3		3	2	**6**	**39**
rank, $\overline{\mathcal{D}}_m$	**7**	**7**	**8**	4	3	**5**		2	**6**	**42**
top, $\hat{\mathcal{D}}_m$	**6**	**7**	**7**	**5**	**5**	**6**	**6**		**7**	**49**
top, $\overline{\mathcal{D}}_m$	4	**7**	**6**	1	2	2	2	1		**25**

Table 18. Wins for $k = 50$ on 3 subfolds

k=50, sf=5	baseline	chi, $\hat{\mathcal{D}}_m$	chi, $\overline{\mathcal{D}}_m$	random, $\hat{\mathcal{D}}_m$	random, $\overline{\mathcal{D}}_m$	rank, $\hat{\mathcal{D}}_m$	rank, $\overline{\mathcal{D}}_m$	top, $\hat{\mathcal{D}}_m$	top, $\overline{\mathcal{D}}_m$	Σ
baseline		**5**	**5**	1	2	0	3	1	2	**19**
chi, $\hat{\mathcal{D}}_m$	3		4	0	2	1	1	1	2	**14**
chi, $\overline{\mathcal{D}}_m$	3	4		1	2	1	3	2	3	**19**
random, $\hat{\mathcal{D}}_m$	**7**	**8**	**7**		**7**	3	**7**	4	**7**	**50**
random, $\overline{\mathcal{D}}_m$	**6**	**6**	**6**	1		1	**6**	3	**5**	**34**
rank, $\hat{\mathcal{D}}_m$	**8**	**7**	**7**	**5**	**7**		**8**	3	**7**	**52**
rank, $\overline{\mathcal{D}}_m$	**5**	**7**	**5**	1	2	0		1	**5**	**26**
top, $\hat{\mathcal{D}}_m$	**7**	**7**	**6**	4	**5**	**5**	**7**		**5**	**46**
top, $\overline{\mathcal{D}}_m$	**6**	**6**	**5**	1	3	1	3	3		**28**

Table 19. Wins for $k = 50$ on 5 subfolds

k=50, sf=7	baseline	chi, $\hat{\mathcal{D}}_m$	chi, $\overline{\mathcal{D}}_m$	random, $\hat{\mathcal{D}}_m$	random, $\overline{\mathcal{D}}_m$	rank, $\hat{\mathcal{D}}_m$	rank, $\overline{\mathcal{D}}_m$	top, $\hat{\mathcal{D}}_m$	top, $\overline{\mathcal{D}}_m$	Σ
baseline		**5**	**7**	0	2	0	1	3	3	**21**
chi, $\hat{\mathcal{D}}_m$	3		**5**	0	1	1	1	2	2	**15**
chi, $\overline{\mathcal{D}}_m$	1	3		1	1	0	1	2	2	**11**
random, $\hat{\mathcal{D}}_m$	**8**	**8**	**7**		**6**	**5**	**7**	**5**	**8**	**54**
random, $\overline{\mathcal{D}}_m$	**6**	**7**	**7**	2		4	**6**	**5**	**6**	**43**
rank, $\hat{\mathcal{D}}_m$	**8**	**7**	**8**	3	4		**7**	**7**	**8**	**52**
rank, $\overline{\mathcal{D}}_m$	**7**	**7**	**7**	1	2	1		2	**6**	**33**
top, $\hat{\mathcal{D}}_m$	**5**	**6**	**6**	3	3	1	**6**		**7**	**37**
top, $\overline{\mathcal{D}}_m$	**5**	**6**	**6**	0	2	0	2	1		**22**

Table 20. Wins for $k = 50$ on 7 subfolds

k=75, sf=3	baseline	chi, $\hat{\mathcal{D}}_m$	chi, $\overline{\mathcal{D}}_m$	random, $\hat{\mathcal{D}}_m$	random, $\overline{\mathcal{D}}_m$	rank, $\hat{\mathcal{D}}_m$	rank, $\overline{\mathcal{D}}_m$	top, $\hat{\mathcal{D}}_m$	top, $\overline{\mathcal{D}}_m$	Σ
baseline		**6**	**7**	2	3	2	3	2	2	**27**
chi, $\hat{\mathcal{D}}_m$	2		**5**	1	1	1	1	1	1	**13**
chi, $\overline{\mathcal{D}}_m$	1	3		0	1	0	1	1	1	**8**
random, $\hat{\mathcal{D}}_m$	**6**	**7**	**8**		4	2	**6**	**5**	**5**	**43**
random, $\overline{\mathcal{D}}_m$	**5**	**7**	**7**	4		4	**5**	4	**5**	**41**
rank, $\hat{\mathcal{D}}_m$	**6**	**7**	**8**	**6**	4		**6**	**5**	**6**	**48**
rank, $\overline{\mathcal{D}}_m$	**5**	**7**	**7**	2	3	2		3	3	**32**
top, $\hat{\mathcal{D}}_m$	**6**	**7**	**7**	3	4	3	**5**		3	**38**
top, $\overline{\mathcal{D}}_m$	**6**	**7**	**7**	3	3	2	**5**	**5**		**38**

Table 21. Wins for $k = 75$ on 3 subfolds

k=75, sf=5	baseline	chi, $\hat{\mathcal{D}}_m$	chi, $\overline{\mathcal{D}}_m$	random, $\hat{\mathcal{D}}_m$	random, $\overline{\mathcal{D}}_m$	rank, $\hat{\mathcal{D}}_m$	rank, $\overline{\mathcal{D}}_m$	top, $\hat{\mathcal{D}}_m$	top, $\overline{\mathcal{D}}_m$	Σ
baseline		**5**	**7**	0	1	0	3	2	3	**21**
chi, $\hat{\mathcal{D}}_m$	3		4	0	2	1	2	1	3	**16**
chi, $\overline{\mathcal{D}}_m$	1	4		0	0	0	1	1	1	**8**
random, $\hat{\mathcal{D}}_m$	**8**	**8**	**8**		**8**	**6**	**7**	**6**	**8**	**59**
random, $\overline{\mathcal{D}}_m$	**7**	**6**	**8**	0		3	3	3	**6**	**36**
rank, $\hat{\mathcal{D}}_m$	**8**	**7**	**8**	2	**5**		**7**	**6**	**8**	**51**
rank, $\overline{\mathcal{D}}_m$	**5**	**6**	**7**	1	**5**	1		1	**6**	**32**
top, $\hat{\mathcal{D}}_m$	**6**	**7**	**7**	2	**5**	2	**7**		**7**	**43**
top, $\overline{\mathcal{D}}_m$	**5**	**5**	**7**	0	2	0	2	1		**22**

Table 22. Wins for $k = 75$ on 5 subfolds

k=75, sf=7	baseline	chi, $\hat{\mathcal{D}}_m$	chi, $\overline{\mathcal{D}}_m$	random, $\hat{\mathcal{D}}_m$	random, $\overline{\mathcal{D}}_m$	rank, $\hat{\mathcal{D}}_m$	rank, $\overline{\mathcal{D}}_m$	top, $\hat{\mathcal{D}}_m$	top, $\overline{\mathcal{D}}_m$	Σ
baseline		**5**	**7**	0	3	1	3	2	4	**25**
chi, $\hat{\mathcal{D}}_m$	3		**5**	2	4	2	3	2	3	**24**
chi, $\overline{\mathcal{D}}_m$	1	3		0	0	0	1	2	1	**8**
random, $\hat{\mathcal{D}}_m$	**8**	**6**	**8**		**7**	**6**	**8**	**5**	**8**	**56**
random, $\overline{\mathcal{D}}_m$	**5**	4	**8**	1		2	**5**	4	**5**	**34**
rank, $\hat{\mathcal{D}}_m$	**7**	**6**	**8**	2	**6**		**7**	**7**	**7**	**50**
rank, $\overline{\mathcal{D}}_m$	**5**	**5**	**7**	0	3	1		2	3	**26**
top, $\hat{\mathcal{D}}_m$	**6**	**6**	**6**	3	4	1	**6**		**5**	**37**
top, $\overline{\mathcal{D}}_m$	4	**5**	**7**	0	3	1	**5**	3		**28**

Table 23. Wins for $k = 75$ on 7 subfolds

k=100, sf=3	baseline	chi, $\hat{\mathcal{D}}_m$	chi, $\overline{\mathcal{D}}_m$	random, $\hat{\mathcal{D}}_m$	random, $\overline{\mathcal{D}}_m$	rank, $\hat{\mathcal{D}}_m$	rank, $\overline{\mathcal{D}}_m$	top, $\hat{\mathcal{D}}_m$	top, $\overline{\mathcal{D}}_m$	Σ
baseline		**7**	**7**	4	**7**	**6**	**8**	**7**	**8**	**54**
chi, $\hat{\mathcal{D}}_m$	1		**6**	2	2	2	2	1	2	**18**
chi, $\overline{\mathcal{D}}_m$	1	2		0	2	2	2	2	3	**14**
random, $\hat{\mathcal{D}}_m$	4	**6**	**8**		**7**	**7**	**7**	**6**	**7**	**52**
random, $\overline{\mathcal{D}}_m$	1	**6**	**6**	1		4	4	4	**6**	**32**
rank, $\hat{\mathcal{D}}_m$	2	**6**	**6**	1	4		4	3	**5**	**31**
rank, $\overline{\mathcal{D}}_m$	0	**6**	**6**	1	4	4		4	4	**29**
top, $\hat{\mathcal{D}}_m$	1	**7**	**6**	2	4	**5**	4		4	**33**
top, $\overline{\mathcal{D}}_m$	0	**6**	**5**	1	2	3	4	4		**25**

Table 24. Wins for $k = 100$ on 3 subfolds

k=100, sf=5	baseline	chi, $\hat{\mathcal{D}}_m$	chi, $\overline{\mathcal{D}}_m$	random, $\hat{\mathcal{D}}_m$	random, $\overline{\mathcal{D}}_m$	rank, $\hat{\mathcal{D}}_m$	rank, $\overline{\mathcal{D}}_m$	top, $\hat{\mathcal{D}}_m$	top, $\overline{\mathcal{D}}_m$	Σ
baseline		**6**	**8**	2	**6**	0	4	4	**7**	**37**
chi, $\hat{\mathcal{D}}_m$	2		**6**	2	3	2	2	2	3	**22**
chi, $\overline{\mathcal{D}}_m$	0	2		0	0	0	1	0	1	**4**
random, $\hat{\mathcal{D}}_m$	**6**	**6**	**8**		**7**	2	**5**	**6**	**7**	**47**
random, $\overline{\mathcal{D}}_m$	2	**5**	**8**	1		1	3	4	4	**28**
rank, $\hat{\mathcal{D}}_m$	**8**	**6**	**8**	**6**	**7**		**8**	**8**	**8**	**59**
rank, $\overline{\mathcal{D}}_m$	4	**6**	**7**	3	**5**	0		3	**6**	**34**
top, $\hat{\mathcal{D}}_m$	4	**6**	**8**	2	4	0	**5**		**7**	**36**
top, $\overline{\mathcal{D}}_m$	1	**5**	**7**	1	4	0	2	1		**21**

Table 25. Wins for $k = 100$ on 5 subfolds

k=100, sf=7	baseline	chi, $\hat{\mathcal{D}}_m$	chi, $\overline{\mathcal{D}}_m$	random, $\hat{\mathcal{D}}_m$	random, $\overline{\mathcal{D}}_m$	rank, $\hat{\mathcal{D}}_m$	rank, $\overline{\mathcal{D}}_m$	top, $\hat{\mathcal{D}}_m$	top, $\overline{\mathcal{D}}_m$	Σ
baseline		**5**	**7**	2	4	3	4	4	**6**	**35**
chi, $\hat{\mathcal{D}}_m$	3		**8**	1	3	2	3	4	**5**	**29**
chi, $\overline{\mathcal{D}}_m$	1	0		0	1	0	1	1	1	**5**
random, $\hat{\mathcal{D}}_m$	**6**	**7**	**8**		**7**	3	**6**	**6**	**7**	**50**
random, $\overline{\mathcal{D}}_m$	4	**5**	**7**	1		1	**6**	**6**	**7**	**37**
rank, $\hat{\mathcal{D}}_m$	**5**	**6**	**8**	**5**	**7**		**7**	**5**	**7**	**50**
rank, $\overline{\mathcal{D}}_m$	4	**5**	**7**	2	2	1		3	**8**	**32**
top, $\hat{\mathcal{D}}_m$	4	4	**7**	2	2	3	**5**		**5**	**32**
top, $\overline{\mathcal{D}}_m$	2	3	**7**	1	1	1	0	3		**18**

Table 26. Wins for $k = 100$ on 7 subfolds