Table 1: Comparison of different single-machine methods with GREEDY and DDISMI using the SVM classifier.

present an empirical evaluation to show that the distributed version of our method (DDISMI) is tens of times faster than its centralized variant (GREEDY). Note that GREEDY itself is as fast as state-of-the-art centralized methods due to its efficient greedy strategy. Then we demonstrate the advantages of the greedy approach over the existing local search method for diversity maximization in terms of performance and running time. After that, we investigate the defined objective function by studying the effect of λ value on the results. Finally, to validate the quality of the objective function, we show a high correlation between the objective value and the classification accuracy on two small-sized datasets. Before elaborating on the empirical results, it should be mentioned that unlike the GREEDY algorithm which arbitrarily selects the first feature, in the implementation, we select the one that has the maximum MI with the class labels vector.

Comparison to the state-of-the-art feature selection methods. In this section, we compare the quality of various centralized feature selection methods with the proposed distributed (DDISMI) and centralized (GREEDY) methods. In order to test the sensitivity of our method to the structure of the dataset, we have used several datasets from a variety of domains with various number of features and instances in addition to the classic datasets in the literature of feature selection. We have described these datasets in detail in the supplemental material.

We considered a variety of MI-based filter methods, namely Mutual Information Quotient (MIQ) (Ding and Peng 2005), Minimum Redundancy Maximum Relevance (mRMR) (Ding and Peng 2005), Joint Mutual Information (JMI) (Yang and Moody 1999), Spectral Conditional Mutual Information (SpecCMI) (Nguyen et al. 2014), Quadratic Programming Feature Selection (QPFS) (Rodriguez-Lujan et al. 2010), and Conditional Mutual Information (CMI)(Cheng et al. 2008) as baselines as well as non MI-based methods, fisher score (Duda, Hart, and Stork 2001) and reliefF (Robnik-Šikonja and Kononenko 2003). Note that prior papers have performed extensive studies (Brown et al. 2012) comparing these methods and we have chosen methods that achieve the best results in various domains. To test the quality of the selected features of each method, we feed them into a classifier method M and compare their classification accuracy. All of the experiments are performed with both SVM and 3-NN as the classifiers (M). Note that all of the methods are filter-based and hence are independent from the selection of M. The LIBSVM package (Chang and Lin 2011) is the underlying implementation of the linear SVM with its regularization factor set to 1. We change $|S|$ from 10 to $min\{100, n\}$, where n is the number of the features in each dataset. Finally, we report the average cross validation (CV) classification accuracy of each method on all of the experiments with different values of $|S|$. A 10-fold CV is being used for the datasets with more than 100 instances and for the others, we employ the leave-one-out CV. In order to compute the probabilities used in MIs and VIs we have discretized the continuous variables using the Minimum Description Length (MDL) method (Irani and Fayyad 1993) with 5 bins. The regularization factor of our algorithm (λ) is set to be 0.8 in all of our experiments. We elaborate on choosing the λ in the experiment about effect of λ value. In order to run DDISMI, we simulate the distributed setting on a single machine and each (simulated) machine only has access to its own feature vectors. Each machine is responsible for processing $\sqrt{nk}$ features when we have n features and want to select k of them. Thus we employ $\sqrt{n/k}$ machines for the first stage. Then, we merge the results of all of the machines together and then process them again, i.e., we select k features from all of these $k\sqrt{n/k} = \sqrt{nk}$ features. For the sake of reproducibility, we have provided all the codes in the supplemental material. Table 1 compares the SVM classification accuracy of the selected features. For each dataset (row), the results with higher accuracy and lower standard deviation are indicated