A Comparative Study of Filter-based Feature Ranking Techniques

Huanjing Wang
Western Kentucky University
Bowling Green, Kentucky 42101
huanjing.wang@wku.edu

Taghi M. Khoshgoftaar
Florida Atlantic University
Boca Raton, Florida 33431
taghi@cse.fau.edu

Kehan Gao
Eastern Connecticut State University
Willimantic, Connecticut 06226
gaok@easternct.edu

***Abstract*—One factor that affects the success of machine learning is the presence of irrelevant or redundant information in the training data set. Filter-based feature ranking techniques (rankers) rank the features according to their relevance to the target attribute and we choose the most relevant features to build classification models subsequently. In order to evaluate the effectiveness of different feature ranking techniques, a commonly used method is to assess the classification performance of models built with the respective selected feature subsets in terms of a given performance metric (e.g., classification accuracy or misclassification rate). Since a given performance metric usually can capture only one specific aspect of the classification performance, it may be unable to evaluate the classification performance from different perspectives. Also, there is no general consensus among researchers and practitioners regarding which performance metrics should be used for evaluating classification performance. In this study, we investigated six filter-based feature ranking techniques and built classification models using five different classifiers. The models were evaluated using eight different performance metrics. All experiments were conducted on four imbalanced data sets from a telecommunications software system. The experimental results demonstrate that the choice of a performance metric may significantly influence the classification evaluation conclusion. For example, one ranker may outperform another when using a given performance metric, but for a different performance metric the results may be reversed. In this study, we have found five distinct patterns when utilizing eight performance metrics to order six feature selection techniques.**

I. Introduction

Developing high-quality software is an important goal for any development team. Software metrics (features) that are collected during the software development process include valuable information about a software project's status, progress, quality, and evolution. Predicting the quality of software modules using software metrics in the early stages of the software development process is very critical. However, not all software metrics are relevant to the class attribute. Feature selection [1] is a process of selecting a subset of relevant features for building learning models. When irrelevant features are eliminated from the original data set, the predictive accuracy of quality models can be improved [2]. The quality models are evaluated based on performance metrics computed after the model-training process. Generally, a given performance metric can reflect a specific aspect of classification performance but cannot cover all the characteristics of it. In addition, the related literature lacks general agreement on which performance metrics should be used for evaluating classification performance [3], [4], [5].

In this empirical study, we investigated six different filter-based feature ranking techniques (rankers), chi-square (CS), information gain (IG), gain ratio (GR), symmetrical uncertainty (SU), and two forms of ReliefF (RF and RFW). In order to evaluate the effectiveness of these methods, we built classification models using five different classifiers on the smaller subsets of selected attributes. The five classifiers used in the study include naïve Bayes (NB), multilayer perceptron (MLP), k-nearest neighbors (KNN), support vector machine (SVM), and logistic regression (LR). Each classification model is assessed with eight different performance metrics: the area under the Receiver Operating Characteristic (ROC) curve (AUC), the area under the Precision-Recall curve (PRC), Default F-Measure (DFM), Best F-Measure (BFM), Default Geometric Mean (DGM), Best Geometric Mean (BGM), Default Arithmetic Mean (DAM), and Best Arithmetic Mean (BAM).

The empirical validation of the different models was implemented through a case study of four imbalanced data sets from a telecommunications software system. Each data set holds the same number of attributes but has a different number of observations. The results demonstrate that the selection of a performance metric may directly impact the evaluation outcome. For instance, one ranker may perform better than another ranker in terms of a given performance metric, but this may not be true when using a different performance metric. In this study, we have discovered five distinct patterns when we used eight performance metrics to order six feature selection techniques.

The main contribution of this work is to provide an assessment and comparison of six filter-based feature ranking techniques using eight performance metrics and over five different classifiers. To our knowledge, no one has done such an extensive study yet.

The rest of the paper is organized as follows. Section II provides more detailed information about the techniques used in the study. The software measurement data sets used in the experiment are described in Section III. Section IV presents the experimental results and analysis. Finally, the conclusion is summarized in Section V.

II. Methodology

A. Filter-based Feature Ranking Techniques

Filter-based feature ranking techniques rank features independently without involving any learning algorithm. Feature ranking consists of scoring each feature according to a particular method, then selecting features based on their scores. This work employs some commonly used filter-based feature ranking techniques including chi-square, information gain, gain ratio, symmetrical uncertainty, and ReliefF. The chi-square (CS) [6] test is used to examine if there is 'no association' between two attributes, i.e., whether the two variables are independent. Information gain, gain ratio, and symmetrical uncertainty are measures based on the concept of entropy, which is based on information theory. Information gain (IG) [7] is the information provided about the target class attribute Y, given the value of independent attribute X. Information gain measures the decrease of the weighted average impurity of the partitions, compared with the impurity of the complete set of data. A drawback of IG is that it tends to prefer attributes with a larger number of possible values. One strategy to counter this problem is to use the gain ratio (GR), which

IEEE IRI 2010, August 4-6, 2010, Las Vegas, Nevada, USA
978-1-4244-8098-2/10/$26.00 ©2010 IEEE

penalizes multi-valued attributes. Symmetrical uncertainty (SU) [8] is another way to overcome the problem of IG's bias toward attributes with more values, doing so by dividing IG by the sum of the entropies of X and Y. Relief is an instance-based feature ranking technique [9]. ReliefF is an extension of the Relief algorithm that can handle noise and multi-class data sets. When the 'weightByDistance' (weight nearest neighbors by their distance) parameter is set as default (false), the algorithm is referred to as RF; when the parameter is set to true, the algorithm is referred to as RFW.

B. Classifiers

Software quality models are built with five well-known classification algorithms [10] including naïve Bayes (NB), multilayer perceptron (MLP), k-nearest neighbors (KNN), support vector machine (SVM) and logistic regression (LR). These were selected because of their common use in software engineering and other data mining applications. Unless stated otherwise, we use default parameter settings for the different learners as specified in the WEKA [10] data mining tool. Parameter settings are changed only when a significant improvement in performance is obtained. For the KNN classifier, 5 neighbors are used in the study.

C. Performance Metrics

In a two-group classification problem, such as fault-prone (positive) and not fault-prone (negative), there can be four possible prediction outcomes: true positive (TP) (i.e., correctly classified positive instances), false positive (FP) (i.e., negative instance classified as positive), true negative (TN) (i.e., correctly classified as negative instance), and false negative (FN) (i.e., positive instance classified as negative). The numbers of cases from the four sets (outcomes) form the basis for several other performance measures that are well known and commonly used for classifier evaluation.

- *Area Under ROC (Receiver Operating Characteristic) Curve* (AUC): has been widely used to measure classification model performance [11]. AUC is a single-value measurement that ranges from 0 to 1. The ROC curve is used to characterize the trade-off between true positive rate ($\frac{|TP|}{|TP|+|FN|}$) and false positive rate ($\frac{|FP|}{|FP|+|TN|}$). A perfect classifier provides an AUC that equals 1.
- *Area Under the Precision-Recall Curve* (PRC): is a single-value measure that originated from the area of information retrieval. The area under the PRC ranges from 0 to 1. The PRC diagram depicts the trade off between recall ($\frac{|TP|}{|TP|+|FN|}$) and precision ($\frac{|TP|}{|TP|+|FP|}$). A classifier that is near optimal in AUC space may not be optimal in precision/recall space.
- *Default F-measure* (DFM): The F-measure is a single value metric that originated from the field of information retrieval [12]. It is calculated as $\frac{2|TP|}{2|TP|+|FP|+|FN|}$. The Default F-measure (DFM) corresponds to a decision threshold value of 0.5.
- *Best F-Measure (BFM)*: is the largest value of F-measure when varying the decision threshold value between 0 and 1. A perfect classifier yields an F-measure of 1, i.e., no misclassification.
- *Default Geometric Mean* (DGM): The Geometric Mean (GM) is a single-value performance measure that ranges from 0 to 1, and a perfect classifier provides a value of 1. GM is defined as the square root of the product of true positive rate and true negative rate, where the true negative rate is defined as $\frac{|TN|}{|FP|+|TN|}$. The decision threshold t = 0.5 is used for the Default Geometric Mean (DGM).

TABLE I
SOFTWARE DATA SETS CHARACTERISTICS

	Data	#Metrics	#Modules	*%fp*	*%nfp*
LLTS	SP1	42	3649	6.28%	93.72%
	SP2	42	3981	4.75%	95.25%
	SP3	42	3541	1.33%	98.67%
	SP4	42	3978	2.31%	97.69%

TABLE II
PERFORMANCE METRICS USING NB

Data	Ranker	AUC	PRC	DFM	BFM	DGM	BGM	DAM	BAM
SP1	CS	0.7846	0.2331	0.2895	0.3045	0.5616	0.7227	0.6356	0.7241
	IG	0.7346	0.216	0.2777	0.2942	0.5113	0.6966	0.6138	0.7018
	GR	0.7831	0.2271	0.294	0.3109	**0.5706**	0.7204	0.6404	0.7220
	RF	0.7879	0.213	0.2706	0.2953	0.5394	0.7301	0.6226	0.7309
	RFW	**0.7882**	0.2145	0.2682	0.2888	0.5397	**0.7320**	0.6222	**0.7326**
	SU	0.7865	**0.2420**	**0.3046**	**0.3140**	0.5676	0.7214	**0.6411**	0.7231
SP2	CS	**0.8108**	0.1975	**0.2797**	0.2915	**0.5891**	**0.7526**	**0.6535**	**0.7532**
	IG	0.7613	**0.1988**	0.2793	**0.2967**	0.5314	0.7217	0.627	0.7241
	GR	0.8081	0.1886	0.2629	0.272	0.5617	0.7524	0.6376	0.7528
	RF	0.8053	0.1941	0.2409	0.2649	0.5353	0.7314	0.622	0.7337
	RFW	0.8081	0.1974	0.242	0.2677	0.5367	0.7335	0.6228	0.7363
	SU	0.7729	0.1806	0.2682	0.2831	0.5511	0.7281	0.6341	0.7295
SP3	CS	0.8184	0.072	0.1319	0.1561	0.5663	0.7689	0.6435	0.7705
	IG	0.7808	0.0603	0.1203	0.1398	0.53	0.7437	0.6234	0.7457
	GR	0.8118	0.0721	**0.1384**	0.1563	**0.5884**	0.7663	**0.6566**	0.7678
	RF	**0.8305**	**0.0767**	0.1285	**0.1608**	0.5435	**0.7952**	0.6316	**0.7957**
	RFW	0.819	0.0744	0.1303	0.1596	0.5492	0.7662	0.6346	0.7688
	SU	0.7882	0.0645	0.1238	0.1467	0.5446	0.7476	0.6311	0.7489
SP4	CS	0.7696	0.1229	0.2094	0.2358	**0.6211**	0.7286	**0.6757**	0.7328
	IG	0.7519	0.1121	**0.2189**	0.2307	0.5798	0.722	0.654	0.7298
	GR	**0.7794**	0.1103	0.1943	0.2098	0.5984	**0.7292**	0.6605	**0.7332**
	RF	0.7731	0.124	0.2146	0.245	0.5967	0.7267	0.6624	0.7295
	RFW	0.7735	**0.1273**	0.2172	**0.2533**	0.6002	0.726	0.6646	0.7286
	SU	0.7592	0.1105	0.2124	0.2257	0.5891	0.7262	0.658	0.7317

- *Best Geometric Mean (BGM)*: is the maximum Geometric Mean value that is obtained when varying the decision threshold between 0 and 1.
- *Default Arithmetic Mean* (DAM): The arithmetic mean is just like the geometric mean but uses the arithmetic mean of the true positive rate and true negative rate instead of the geometric mean. It is also a single-value performance measure that ranges from 0 to 1. The decision threshold t = 0.5 is used for the Default Arithmetic Mean (DAM).
- *Best Arithmetic Mean* (BAM): is just like the BGM, but using the maximum arithmetic mean that is obtained when varying the decision threshold between 0 and 1.

III. DATA SET CHARACTERISTICS

Experiments conducted in this study used software metrics and defect data collected from a real-world software project, a very large telecommunications software system (denoted as LLTS) [13]. LLTS contains data from four consecutive releases, which are labeled as SP1, SP2, SP3, and SP4. The software measurement data sets consist of 42 software metrics, including 24 product metrics, 14 process metrics, and four execution metrics [13]. The dependent variable is the class of the program module, fault-prone (*fp*), or not fault-prone (*nfp*). A program module with one or more faults is considered *fp*, and *nfp* otherwise. Table I lists the characteristics of the four release data sets utilized in this work. An important characteristic of these data sets is that they all suffer from class imbalance, where the proportion of *fp* modules is much lower than that of *nfp* modules.

IV. EXPERIMENTS

A. Experimental Design

We first used six filter-based rankers to select the subsets of attributes. We ranked the features and selected the top $\lceil \log_2 n \rceil$ features according to their respective scores, where n is the number of independent features for a given data set. The reasons why we select the top $\lceil \log_2 n \rceil$ features include (1) related literature does not provide guidance on the appropriate number of features to select; and

TABLE III
PERFORMANCE METRICS USING MLP

Data	Ranker	AUC	PRC	DFM	BFM	DGM	BGM	DAM	BAM
SP1	CS	0.7943	0.2688	0.128	**0.3389**	0.2682	0.7318	0.5343	0.7330
	IG	0.7475	0.2397	0.1309	0.297	0.272	0.7003	0.5352	0.7037
	GR	0.7926	0.2682	0.1294	0.3331	0.2696	0.7315	0.5348	0.7325
	RF	0.7948	0.2504	0.062	0.3286	0.18	0.7372	0.5154	0.7382
	RFW	**0.7955**	0.2513	0.0532	0.3241	0.167	**0.7384**	0.5129	**0.7390**
	SU	0.7875	**0.2754**	**0.1562**	0.3237	**0.2988**	0.7262	**0.5430**	0.7274
SP2	CS	0.8117	**0.2423**	0.0863	**0.3147**	0.2128	**0.7456**	0.5226	**0.7467**
	IG	0.7545	0.2033	0.0936	0.2761	0.2266	0.7106	0.5246	0.7129
	GR	0.8099	0.2298	0.0614	0.2944	0.1759	0.7418	0.5157	0.7437
	RF	0.8119	0.2415	**0.1088**	0.3042	**0.2435**	0.7422	**0.5290**	0.7434
	RFW	**0.8139**	0.2413	0.1062	0.3049	0.2403	0.7411	0.5283	0.7433
	SU	0.7847	0.2061	0.1007	0.2723	0.2347	0.7262	0.5268	0.7277
SP3	CS	0.8126	0.0861	**0.0041**	0.1919	**0.0146**	0.7565	**0.5010**	0.7600
	IG	0.7688	0.0642	0	0.1463	0	0.7302	0.5	0.7354
	GR	0.8209	**0.0903**	0	**0.2002**	0	0.7621	0.5	0.7646
	RF	0.8191	0.0842	0	0.1887	0	0.7655	0.5	0.7672
	RFW	**0.8303**	0.0885	0	0.1944	0	**0.7726**	0.5	**0.7737**
	SU	0.7843	0.0714	0	0.1621	0	0.7369	0.5	0.7392
SP4	CS	0.7914	0.1367	0.0221	0.2459	0.0831	0.7308	0.5053	0.7344
	IG	0.7464	0.1191	0.0105	0.2292	0.0521	0.7101	0.5024	0.7170
	GR	**0.8103**	0.1444	0.022	0.2505	0.0821	**0.7472**	0.5054	**0.7498**
	RF	0.7619	**0.1533**	0.0285	**0.2693**	0.1073	0.7063	0.5072	0.7087
	RFW	0.7598	0.1529	**0.0322**	0.2687	**0.1215**	0.7029	**0.5081**	0.7066
	SU	0.7504	0.1202	0.0081	0.233	0.0285	0.7093	0.5017	0.7165

TABLE IV
PERFORMANCE METRICS USING KNN

Data	Ranker	AUC	PRC	DFM	BFM	DGM	BGM	DAM	BAM
SP1	CS	0.757	**0.1880**	**0.1392**	**0.2544**	**0.2958**	**0.7122**	**0.5382**	**0.7138**
	IG	0.7139	0.1632	0.1276	0.2404	0.284	0.6595	0.5343	0.6679
	GR	0.7475	0.176	0.1295	0.2423	0.2833	0.7017	0.5349	0.7033
	RF	0.7495	0.1735	0.1169	0.2502	0.2735	0.7034	0.5304	0.7046
	RFW	0.7489	0.1655	0.1029	0.247	0.2547	0.7045	0.5256	0.7061
	SU	**0.7600**	0.1774	0.1279	0.2482	0.2842	0.7082	0.5343	0.7093
SP2	CS	**0.7800**	0.1588	0.1017	0.2165	0.2457	0.7221	0.5271	0.7259
	IG	0.7515	0.1495	0.1084	0.2288	0.2564	0.6971	0.5294	0.7026
	GR	0.7721	0.1379	0.0799	0.2029	0.2164	**0.7244**	0.5203	**0.7292**
	RF	0.7221	0.1181	0.0595	0.1986	0.187	0.6781	0.5137	0.6796
	RFW	0.7255	0.1175	0.0526	0.1963	0.1753	0.6801	0.5115	0.6819
	SU	0.7796	**0.1680**	**0.1233**	**0.2299**	**0.2724**	0.7205	**0.5342**	0.7231
SP3	CS	0.7879	0.0549	0	0.1184	0	**0.7415**	0.4996	**0.7438**
	IG	0.7298	0.0433	0	0.1015	0	0.7028	0.4997	0.7042
	GR	0.7802	**0.0552**	0	0.1227	0	0.7392	0.4995	0.7410
	RF	**0.7898**	0.049	0	0.1113	0	0.7383	0.4996	0.7406
	RFW	0.7704	0.0539	0	**0.1255**	0	0.7303	0.4991	0.7323
	SU	0.7602	0.0536	**0.0078**	0.1086	**0.0292**	0.7279	**0.5017**	0.7296
SP4	CS	0.7913	0.1044	0.0642	**0.1782**	0.1896	**0.7382**	0.5174	**0.7400**
	IG	0.6853	0.0751	0.0438	0.1485	0.1546	0.6577	0.5111	0.6687
	GR	**0.7967**	**0.1117**	0.068	0.1781	0.1919	0.7363	0.5183	0.7381
	RF	0.7631	0.0904	0.0736	0.1501	0.2076	0.7268	0.5204	0.7276
	RFW	0.7665	0.0922	**0.0752**	0.1534	**0.2098**	0.7259	**0.5209**	0.7266
	SU	0.7433	0.0815	0.039	0.1535	0.1394	0.7022	0.5098	0.7036

TABLE V
PERFORMANCE METRICS USING SVM

Data	Ranker	AUC	PRC	DFM	BFM	DGM	BGM	DAM	BAM
SP1	CS	0.6401	0.1545	0.0309	0.2341	0.1185	0.6154	0.5073	0.6262
	IG	0.6532	0.1923	0.0706	0.2602	0.192	0.6264	0.5179	0.6447
	GR	0.6651	0.1716	0.0563	0.2443	0.1661	**0.6330**	0.5141	0.6466
	RF	**0.6708**	0.1472	0.0219	0.217	0.0942	0.6244	0.5051	0.6318
	RFW	0.6368	0.1313	0.0136	0.1891	0.0752	0.5979	0.5032	0.6036
	SU	0.6632	**0.1992**	**0.0847**	**0.2678**	**0.2079**	0.6258	**0.5219**	**0.6476**
SP2	CS	**0.7060**	**0.1526**	0.0492	0.2303	0.1565	**0.6645**	0.5123	**0.6718**
	IG	0.6577	0.1522	**0.0528**	**0.2434**	**0.1643**	0.6356	**0.5132**	0.6567
	GR	0.6628	0.1352	0.0253	0.2125	0.1058	0.6343	0.506	0.6465
	RF	0.6357	0.1121	0.0194	0.1742	0.0801	0.6016	0.5048	0.6101
	RFW	0.6386	0.1211	0.0124	0.1808	0.0592	0.5949	0.5031	0.6060
	SU	0.6572	0.1435	0.0484	0.2202	0.1507	0.6363	0.512	0.6508
SP3	CS	0.6456	0.0448	0	0.1143	0	0.6351	0.4999	0.6514
	IG	0.6294	0.0379	0	0.1005	0	0.6219	0.4999	0.6373
	GR	0.647	0.0417	0	0.1027	0	0.6292	0.4998	0.6516
	RF	0.6341	0.0352	**0.0038**	0.092	**0.0146**	0.6162	**0.5009**	0.6274
	RFW	**0.6611**	**0.0470**	0	**0.1296**	0	0.6349	0.4998	0.6432
	SU	0.6601	0.047	0	0.1182	0	**0.6529**	0.4998	**0.6624**
SP4	CS	0.6529	0.071	**0.0145**	0.1421	**0.0668**	0.6334	**0.5034**	0.6476
	IG	0.6247	0.0736	0.0124	0.1528	0.0564	0.6227	0.5028	0.6437
	GR	**0.6531**	**0.0741**	0.0042	0.1457	0.0208	**0.6406**	0.5008	**0.6497**
	RF	0.6248	0.0604	0.0021	0.1334	0.0104	0.6021	0.5004	0.6165
	RFW	0.6423	0.074	0.0021	**0.1584**	0.0104	0.6188	0.5005	0.6318
	SU	0.6399	0.0687	0.0041	0.1455	0.0208	0.6244	0.5007	0.6411

TABLE VI
PERFORMANCE METRICS USING LR

Data	Ranker	AUC	PRC	DFM	BFM	DGM	BGM	DAM	BAM
SP1	CS	0.8021	0.276	0.1412	**0.3449**	0.2862	0.7363	0.5384	0.7374
	IG	0.7688	0.2574	0.1554	0.3069	0.2999	0.7196	0.5429	0.7204
	GR	0.8014	0.2734	0.1453	0.3374	0.2901	0.7353	0.5396	0.7366
	RF	**0.8103**	0.2652	0.1045	0.3229	0.2412	**0.7486**	0.5272	**0.7492**
	RFW	0.8091	0.2661	0.1129	0.3207	0.2509	0.7475	0.5297	0.7479
	SU	0.7993	**0.2854**	**0.1748**	0.3384	**0.3207**	0.738	**0.5490**	0.7388
SP2	CS	0.8229	**0.2579**	**0.1403**	**0.3156**	**0.2856**	0.7493	**0.5389**	0.7504
	IG	0.7935	0.2191	0.1222	0.2895	0.2654	0.7325	0.5334	0.7339
	GR	0.8176	0.238	0.1164	0.3001	0.2553	0.7429	0.5313	0.7436
	RF	0.8221	0.246	0.1123	0.3139	0.2505	0.7523	0.5301	0.7531
	RFW	**0.8233**	0.247	0.1186	0.3154	0.2581	**0.7538**	0.532	**0.7547**
	SU	0.7909	0.2144	0.1197	0.2799	0.2621	0.7322	0.5325	0.7342
SP3	CS	0.8354	0.0866	0	0.1829	0	0.7844	0.4993	0.7856
	IG	0.7805	0.0656	0	0.1537	0	0.7425	0.4993	0.7448
	GR	0.8361	**0.0880**	0.0038	0.1832	0.0146	0.7881	0.5004	0.7886
	RF	0.8354	0.0874	0.0038	**0.1931**	0.0146	0.7836	0.5005	0.7845
	RFW	**0.8387**	0.085	**0.0039**	0.1842	**0.0146**	**0.7940**	**0.5007**	**0.7950**
	SU	0.804	0.0744	0.0038	0.1654	0.0146	0.7561	0.5004	0.7584
SP4	CS	0.8153	0.1449	0.0481	0.248	0.1604	0.7492	0.5124	0.7500
	IG	0.7816	0.1322	**0.0636**	0.2316	**0.1874**	0.733	**0.5169**	0.7359
	GR	**0.8216**	0.1432	0.0463	0.2536	0.158	**0.7556**	0.5119	**0.7566**
	RF	0.8118	0.1517	0.0386	0.2571	0.1443	0.7474	0.5097	0.7481
	RFW	0.8142	**0.1555**	0.0405	**0.2638**	0.1486	0.7477	0.5102	0.7485
	SU	0.7802	0.1247	0.05	0.2278	0.1637	0.726	0.513	0.7281

(2) a recent study [14] showed that it was appropriate to use $\lceil \log_2 n \rceil$ as the number of features when using WEKA [10] to build Random Forests learners for binary classification in general and imbalanced data sets in particular. Although we used different learners here, a preliminary study showed that $\lceil \log_2 n \rceil$ is still appropriate for various learners. In this study, six ($\lceil \log_2 42 \rceil = 6$) features are selected.

The experiments were conducted to discover the impact of (1) eight different performance metrics; (2) six commonly used filter-based rankers; and (3) five different learners. In the experiments, ten runs of five-fold cross-validation were performed. In total, 6,000 models were evaluated during our experiments.

B. Experimental Results

The classification models were evaluated in terms of the eight performance metrics. All the results are reported in Table II through Table VI. Note that each value presented in the table is the average over the ten runs of five-fold cross-validation outcomes. The best model for each data set is indicated in **boldfaced** print. A total of 960 values are included in the five tables. It has been noted that some performance results on SP3 in terms of DFM and DGM for MLP, KNN, SVM and LR learners are zeros since the true positive rate of the corresponding models are zeros. From these tables, we can observe that when one ranker performed best in terms of one performance metric, this may not be true when other performance metrics are used to evaluate models. For example, RFW performed best in terms of AUC, SU performed best in terms of PRC, GR performed best on performance metric DGM when models are built using the SP1 data set and NB classifier (see Table II), CS performed best in terms of BFM performance metric, and RF performed best in terms of BGM when models are built using the SP1 data set and LR classifier (see Table VI). Fig. 1 presents the number of winners for all the rankers in terms of each performance metric. Each result is summarized across all five classifiers and four data sets together.

C. Results Analysis

A two-way ANOVA [15] was performed for each of the eight performance metrics (AUC, PRC, DFM, BFM, DGM, BGM, DAM, and BAM) separately. The two factors are Factor A, in which six

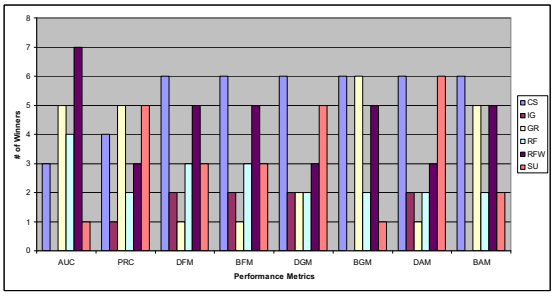


Fig. 1. Summary of Ranker's Performance

TABLE VII
ANALYSIS OF VARIANCE

(a) AUC

Source	Sum Sq.	d.f.	Mean Sq.	F	p-value
A	0.2329	5	0.0466	53.78	0
B	3.8618	4	0.9655	1114.88	0
A×B	0.0684	20	0.0034	3.95	0
Error	1.0132	1170	0.0009		
Total	5.1763	1199			

(b) PRC

Source	Sum Sq.	d.f.	Mean Sq.	F	p-value
A	0.0223	5	0.0045	1.08	0.370
B	1.2106	4	0.3027	73.28	0
A×B	0.0583	20	0.0029	0.71	0.823
Error	4.8325	1170	0.0041		
Total	6.1237	1199			

(c) DFM

Source	Sum Sq.	d.f.	Mean Sq.	F	p-value
A	0.0602	5	0.0120	4.39	0.001
B	5.6876	4	1.4219	519.19	0
A×B	0.0085	20	0.0004	0.16	1.000
Error	3.2043	1170	0.0027		
Total	8.9606	1199			

(d) BFM

Source	Sum Sq.	d.f.	Mean Sq.	F	p-value
A	0.0240	5	0.0048	1.41	0.216
B	1.7631	4	0.4408	129.66	0
A×B	0.0975	20	0.0049	1.43	0.097
Error	3.9773	1170	0.0034		
Total	5.8619	1199			

(e) DGM

Source	Sum Sq.	d.f.	Mean Sq.	F	p-value
A	0.1137	5	0.0227	2.55	0.026
B	35.8865	4	8.9716	1006.31	0
A×B	0.1282	20	0.0064	0.72	0.809
Error	10.4310	1170	0.0089		
Total	46.5594	1199			

(f) BGM

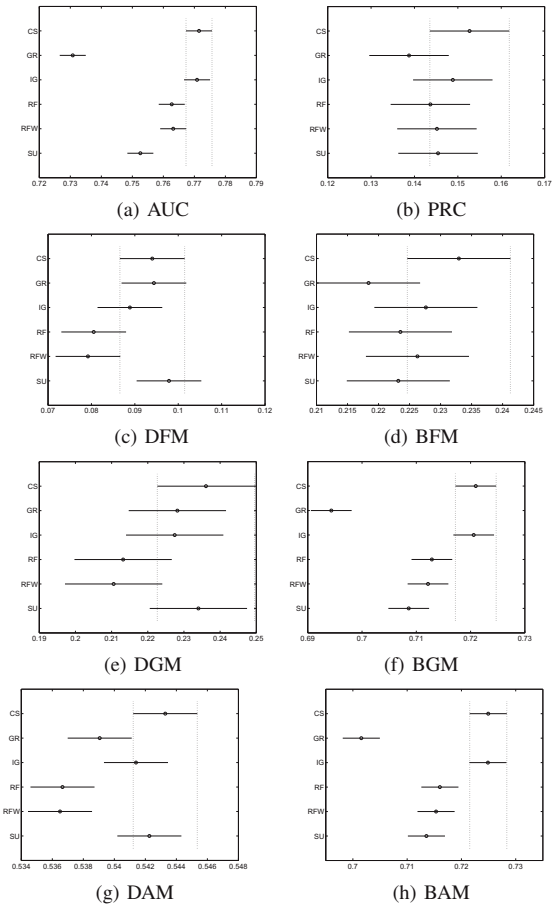
Source	Sum Sq.	d.f.	Mean Sq.	F	p-value
A	0.0954	5	0.0191	27.51	0
B	2.3820	4	0.5955	858.74	0
A×B	0.0699	20	0.0035	5.04	0
Error	0.8114	1170	0.0007		
Total	3.3587	1199			

(g) DAM

Source	Sum Sq.	d.f.	Mean Sq.	F	p-value
A	0.0084	5	0.0017	8.01	0
B	3.0634	4	0.7659	3645.18	0
A×B	0.0118	20	0.0006	2.81	0
Error	0.2458	1170	0.0002		
Total	3.3294	1199			

(h) BAM

Source	Sum Sq.	d.f.	Mean Sq.	F	p-value
A	0.0748	5	0.0150	26.00	0
B	1.9306	4	0.4827	838.88	0
A×B	0.0824	20	0.0041	7.16	0
Error	0.6732	1170	0.0006		
Total	2.7609	1199			



(a) AUC (b) PRC (c) DFM (d) BFM (e) DGM (f) BGM (g) DAM (h) BAM

Fig. 2. Multiple comparisons for Factor A

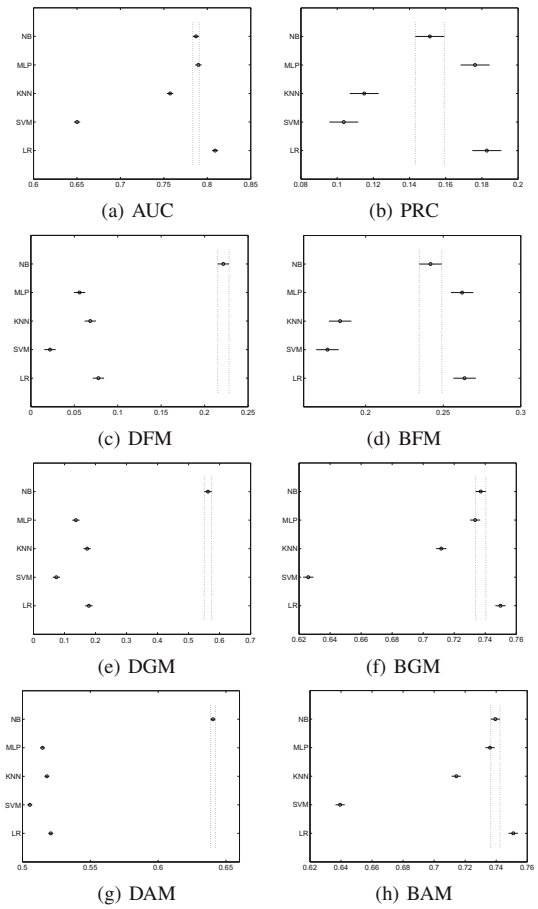
filter-based rankers were considered, and Factor B, in which five classifiers were included. In addition, the interaction A×B was also included. In this ANOVA test, the results from all four release data sets were taken into account together. A significance level of $\alpha = 5\%$ was used for all statistical tests.

The ANOVA results are presented in Table VII. From the table, we can see that for the performance metrics AUC, BGM, DAM and BAM, the p-values for the main factors A and B, and the interaction term A×B were zeros, indicating the performance values are not same for all groups in each of the main factors and also influenced by the interaction term A×B, i.e., Factor A is different at every level of Factor B, and vice versa. For the performance metrics PRC and BFM, there was no significant distinction between any pair of the group means for Factor A and interaction A×B since the p-values were greater than 0.05, while an obvious difference existed in at least one pair of group means for Factor B, because the p-value was zero. For the performance metrics DFM and DGM, an obvious difference existed in at least one pair of group means for Factor A and also for Factor B. However, their interaction did not contribute too much for the classification performance.

Additional multiple comparisons for the main factors and interaction term were performed to investigate the differences among the respective groups (levels). Both ANOVA and multiple comparison tests were implemented in MATLAB. The multiple comparisons are presented in Fig. 2 through 4. The performance of filter-based rankers was ranked from best to worst for each performance metric as shown in Table VIII. Each ranker is labeled with a superscript. The rankers labeled with the same superscripts implies that they were from same performance group, in which no statistically significant difference was found between rankers. The table shows five distinct groups of results when we order six commonly used rankers based on eight performance metrics (over all the classifiers built): (1) PRC, DGM, and BFM (when using these three metrics to evaluate the rankers, the orders of the six rankers are the same or similar.); (2) BGM and BAM (identical ordering of six feature-based rankers); (3) AUC; (4) DFM; and (5) DAM. The performance of learners was also ranked from best to worst for each performance metric as shown in Table IX. We can observe that three distinct patterns emerge when we are ordering learners based on eight performance metrics: (1) AUC, BGM, and BAM; (2) PRC and BFM; and (3) DFM, DGM, and DAM. All the ranks of interaction of rankers and learners are also summarized but not presented here due to space limitations.

Some findings can be summarized from these tables and figures.

- For all performance metrics, there are no significant differences



(a) AUC (b) PRC

(c) DFM (d) BFM

(e) DGM (f) BGM

(g) DAM (h) BAM

Fig. 3. Multiple comparisons for Factor B

TABLE VIII
RANK OF RANKERS (FACTOR **A**)

Ranker	Ranks (best $\longrightarrow$ worst)					
AUC	CS^{a}	IG^{ab}	RFW^{ab}	RF^{b}	SU^{c}	GR^{d}
PRC	CS^{a}	IG^{a}	SU^{a}	RFW^{a}	RF^{a}	GR^{a}
DFM	SU^{a}	GR^{ab}	CS^{abc}	IG^{abc}	RF^{bc}	RFW^{c}
BFM	CS^{a}	IG^{a}	RFW^{a}	RF^{a}	SU^{a}	GR^{a}
DGM	CS^{a}	SU^{a}	GR^{a}	IG^{a}	RF^{a}	RFW^{a}
BGM	CS^{a}	IG^{a}	RF^{b}	RFW^{b}	SU^{b}	GR^{c}
DAM	CS^{a}	SU^{ab}	IG^{ab}	GR^{bc}	RF^{c}	RFW^{c}
BAM	CS^{a}	IG^{a}	RF^{b}	RFW^{b}	SU^{b}	GR^{c}

TABLE IX
RANK OF CLASSIFIERS (FACTOR **B**)

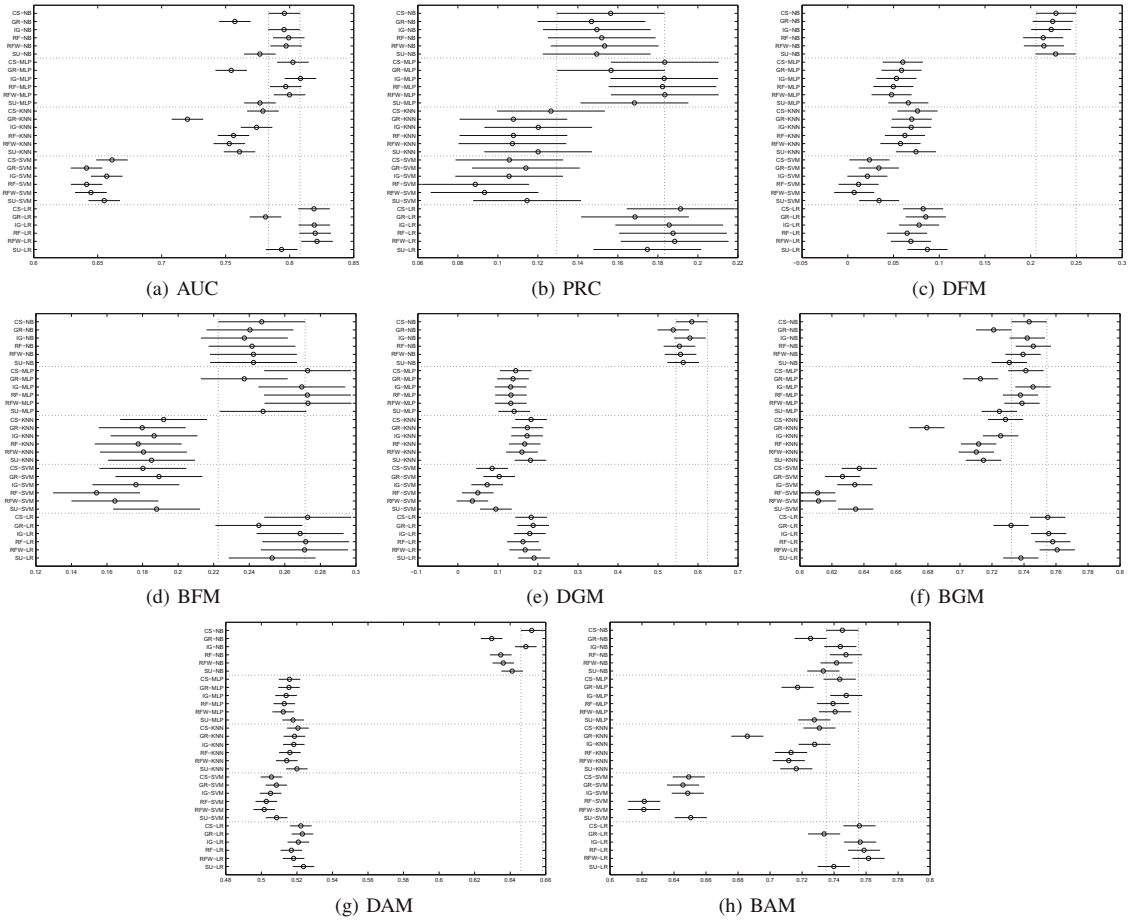
Classifier	Ranks (best $\longrightarrow$ worst)				
AUC	LR^{a}	MLP^{b}	NB^{b}	KNN^{c}	SVM^{d}
PRC	LR^{a}	MLP^{a}	NB^{b}	KNN^{c}	SVM^{c}
DFM	NB^{a}	LR^{b}	KNN^{bc}	MLP^{c}	SVM^{d}
BFM	LR^{a}	MLP^{a}	NB^{b}	KNN^{c}	SVM^{c}
DGM	NB^{a}	LR^{b}	KNN^{b}	MLP^{c}	SVM^{d}
BGM	LR^{a}	NB^{b}	MLP^{b}	KNN^{c}	SVM^{d}
DAM	NB^{a}	LR^{b}	KNN^{bc}	MLP^{c}	SVM^{d}
BAM	LR^{a}	NB^{b}	MLP^{b}	KNN^{c}	SVM^{d}

between CS and IG, the performance differences between RF and RFW are minimal.

- There are no significant differences when ordering all rankers in terms of PRC, DGM, and BFM performance metrics.
- One method being ranked at top by a given performance metric does not mean that it is also ranked at top by another performance metric, and the same for being ranked worst. For example, GR performed worse than other filter-based rankers when using AUC to evaluate classification performance (see Fig. 2(a)), while this is not true when using a different performance metric, for instance, DFM (see Fig. 2(c)).
- CS has the best performance according to all performance metrics except DFM, while SU has the best performance for DFM.
- The performance of various ranking techniques and learners shows two different patterns. One pattern is found when AUC, PRC, BFM, BGM and BAM are utilized for assessment. For Factor A (see Fig. 2), CS performed best, followed by IG; GR performed worst among the six filter-based feature ranking techniques; and RF, RFW and SU sat in between. For Factor B (see Fig. 3), LR performed best, followed by MLP and NB, then KNN, and finally SVM. The other patten appears when DFM, DGM and DAM are used for evaluation. The pattern is that, for Factor A, RF and RFW performed worse than the other four ranking techniques; for factor B, NB significantly outperformed all other learners, followed by LR, KNN, and MLP, and finally SVM. These two patterns are also extended to interaction A$\times$B. The two distinct patterns can be easily observed from Fig. 4.
- The performance distributions of the 30 group means are very similar when evaluated using DFM, DGM and DAM (see Fig. 4(c), 4(e) and 4(g)). The NB group performed much better than the other groups, while the performances of the remaining four groups are relatively close to each other. But still we can see that the KNN and LR groups performed better than MLP and SVM groups. Of the two inferior performance groups, SVM performed even worse. Meanwhile, the performance distributions of the 30 group means show a similar pattern when evaluated in terms of AUC, PRC, BFM, BGM and BAM (see Fig. 4(a), 4(b), 4(d), 4(f) and 4(h)). Overall, the NB, MLP and LR groups present relatively similar performances, but still we can see that the LR group performed best. These three groups outperformed the KNN and SVM groups. In fact, the SVM performed once again worst among the five learner groups. Also, one point that needs to be noted is that if we have to compare the impacts of learners and filter ranking techniques on the classification performance, we can clearly see that learners had more influence on the classification performance in this study.

V. CONCLUSION

In this paper, we present six filter-based feature ranking techniques and evaluate their effectiveness by building five different types of classification models. Each model is assessed in terms of eight performance metrics. The experiments were conducted on four consecutive releases of a very large telecommunications system. The experimental results demonstrate that the selection of a performance metric is critical for assessing classification performance. Using different performance metrics may generate different evaluation results. We summarized five distinct patterns of the six feature ranking techniques when using the eight performance metrics. Every metric concurred on the identification of the worst learner, SVM. These results accentuate



(a) AUC (b) PRC (c) DFM

(d) BFM (e) DGM (f) BGM

(g) DAM (h) BAM

Fig. 4. Multiple comparisons for Factor A×B

the importance of metric selection for learning from class imbalanced data.

More investigations of characteristics of performance metrics and their impact on classification performance using a variety of domain data will be studied in our future work.

REFERENCES

[1] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, March 2003.

[2] H. Wang, T. M. Khoshgoftaar, K. Gao, and N. Seliya, "High-dimensional software engineering data and feature selection," in *Proceedings of 21st IEEE International Conference on Tools with Artificial Intelligence*, Newark, NJ, USA, Nov. 2-5 2009, pp. 83–90.

[3] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation," *AI 2006: Advances in Artificial Intelligence*, pp. 1015–1021, 2006.

[4] N. Seliya, T. M. Khoshgoftaar, and J. Van Hulse, "A study on the relationships of classifier performance metrics," in *Proceedings of the 21st IEEE International Conference on Tools with Artificial Intelligence*, 2009, pp. 59–66.

[5] A. Folleco, T. M. Khoshgoftaar, and A. Napolitano, "Comparison of four performance metrics for evaluating sampling techniques for low quality class-imbalanced data," in *Proceedings of the 2008 Seventh International Conference on Machine Learning and Applications*, Washington, DC, USA, 2008, pp. 153–158.

[6] R. L. Placket, "Karl pearson and the chi-squared test," *International Statistical Review*, vol. 51, no. 1, pp. 59–72, 1983.

[7] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[8] M. A. Hall and G. Holmes, "Benchmarking attribute selection techniques for discrete class data mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 6, pp. 1437 – 1447, Nov/Dec 2003.

[9] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proceedings of 9th International Workshop on Machine Learning*, 1992, pp. 249–256.

[10] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. Morgan Kaufmann, 2005.

[11] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, June 2006.

[12] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, Pennsylvania, 2006, pp. 233–240.

[13] K. Gao, T. M. Khoshgoftaar, and H. Wang, "An empirical investigation of filter attribute selection techniques for software quality classification," in *Proceedings of the 10th IEEE International Conference on Information Reuse and Integration*, Las Vegas, Nevada, August 10-12 2009, pp. 272–277.

[14] T. M. Khoshgoftaar, M. Golawala, and J. Van Hulse, "An empirical study of learning from imbalanced data using random forest," in *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence*, vol. 2, Washington, DC, USA, 2007, pp. 310–317.

[15] M. L. Berenson, M. Goldstein, and D. Levine, *Intermediate Statistical Methods and Applications: A Computer Package Approach*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1983.