Robustness of Filter-Based Feature Ranking: A Case Study

Wilker Altidor and **Taghi M. Khoshgoftaar** and **Jason Van Hulse**
Florida Atlantic University
777 Glades Rd., Boca Raton, FL 33431
wilker.altidor@gmail.com; taghi@cse.fau.edu; jason@gmail.com

Abstract

The filter model of feature selection has been well studied. In previous studies, classification performance has traditionally been proposed as a way to evaluate filter solutions. In this study, a new method of comparing feature ranking techniques is presented providing a straightforward approach for quantifying individual filters' robustness to class noise. Six commonly-used filters, plus one which is rarely used, are investigated regarding their ability to retain, in the presence of class noise, strong classification performance. Three classifiers and one classification performance metric are considered. The experimental results of this study show that Gain Ratio, one of the well known and widely used filters, is very sensitive to class noise. ReliefF offers the best results with both the NB and kNN learners while Signal-to-noise, though not as widely used in the literature as the others, outperforms all the filters with the SVM learner.

Introduction

Feature selection is a pre-processing technique that finds a minimum subset of features that captures the relevant properties of a dataset to enable adequate classification (Gilad-Bachrach, Navot, & Tishby 2004). Given that no loss of relevant information is incurred with a reduction in the original feature space, feature selection has been widely used. Feature selection has been considered in many classification problems (Wang, Khoshgoftaar, & Gao 2010), and it has been used in various application domains (Liu, Li, & Wong 2002) (Ruiz *et al.* 2005). Feature selection techniques are very useful for improving the performance of learning algorithms (Hall & Holmes 2003). For this reason, the strengths and weaknesses of feature selection techniques are traditionally assessed in terms of the classification performance from models built with a subset of the original features.

One of the factors that can characterize real world data is noise introduced in either the independent features, the class labels, or both. Noise in the independent features is referred to as attribute noise while noise in the class labels is called class noise. It is universally accepted that noisy data can be detrimental to data mining and knowledge discovery (Xiong *et al.* 2006). Surprisingly, to the best of our knowledge, there are no studies where noise is taken into consideration when comparing feature ranking techniques, or any other feature selection method. Given the pervasiveness of noise in real world data, understanding the behavior of filters in the presence of noise is of utmost importance. Consequently, a comparison method that is based on both the application of the filters to noisy data and the evaluation of the selected features in terms of classification performances is invaluable. This paper presents such a comparison method while focusing on the filter approach to feature selection. More specifically, this study considers the feature ranking method, which ranks the features from the most to the least relevant (wrapper-based feature ranking methods are not considered here for space considerations, but could be analyzed in future work). Six commonly-used feature ranking techniques (Chi-squared, Information Gain, Gain Ratio, two versions of ReliefF and Symmetric Uncertainty), plus one rarely-used (Signal-to-noise), are studied and compared regarding their ability to maintain, in the presence of class noise, adequate classification performance.

Hence, a new method for comparing filters in terms of their robustness to class noise is presented in this paper. This method involves applying the filters to noisy data, building and testing classification models on noisy and clean data respectively, and comparing the models' classification performances. Three learning algorithms are chosen for this study: Naïve Bayes (NB), k-Nearest Neighbor (kNN), and Support Vector Machine (SVM). The performance of each learning algorithm is determined by the area under the receiver operating characteristic curve metric. Particularly, this study evaluates the effectiveness of seven filters in terms of classification performances at different class noise injection levels. The robustness of each filter is measured by quantifying the variation in the classification performances from models built with the corresponding feature selection. To the best of our knowledge, this paper is the first to present this method of comparing feature ranking techniques, and it shows how a feature ranking technique can be measured in terms of its robustness to class noise.

Related Work

The filter approach to feature selection evaluates feature relevance by examining the intrinsic characteristics of the data

without the use of a classifier (Saeys, Inza, & Larrañaga 2007). Filter-based feature ranking in particular has been well studied. Due to its simplicity, scalability and good empirical success, feature ranking is very attractive (Guyon & Elisseeff 2003).

In many studies, the feature ranking techniques are often compared in terms of the classification performance derived from a subset of the original features. For instance, five feature ranking methods (Document Frequency, Information Gain, Mutual Information, Chi-test and Term Strength) have been evaluated and compared in the study of Yang and Pedersen (Yang & Pedersen 1997). Recall and precision were the two classification performance measures used in their evaluation. Also, in the study of Liu et al. (Liu, Li, & Wong 2002), five selection methods, including four ranking techniques, were considered. The classification performance in terms of accuracy was used to evaluate and compare each feature ranking technique. Moreover, in Méndez et al. (Méndez *et al.* 2006), four feature ranking techniques (Information Gain, Mutual Information, Chi-test, and Document Frequency) were analyzed in terms of their strengths and weaknesses. Six performance metrics were used in the assessment: Overall Accuracy, False Positive Rate, False Negative Rate, Recall, Precision and Total Cost Ratio.

Some works have proposed comparison methods that combined the performance of the learning algorithms with some other schemes. For instance, Hall and Holmes (Hall & Holmes 2003) presented a benchmark comparison of several attribute selection methods for supervised classification. While their study also used classification performance in their evaluation, they presented a "wins versus losses" scheme that consisted of a pairwise comparison between one feature selection technique and another. Ruiz et al. (Ruiz *et al.* 2005), in their analysis of feature rankings, proposed a method based on the Area Under Feature Ranking Classification Curve for comparing different feature ranking techniques. The quality of a feature ranking technique was then determined by the classification performances on three classifiers and the number of times that particular ranking method holds the first position.

This study also uses the classification performance from models built with a subset of the original features to compare different ranking techniques. However, the assessment is done in the presence of class noise, which to our knowledge has not been explored previously. Given the pervasiveness of noise in real world data and the negative effects of noise on data mining and machine learning algorithms, it is imperative to understand the effectiveness of feature selection in the presence of low quality data. This paper shows how to separate the most from the least effective feature ranking techniques and points out the importance of considering the impact of noise on feature selection.

Feature Ranking Techniques

The seven filter-based feature ranking techniques being compared are described below. The first six are commonly used in the literature (chi-squared statistic (χ^2), Information Gain (IG), Gain Ratio (GR), two versions of ReliefF (RF and RFW) and Symmetric Uncertainty (SU)), while the last, Signal-to-noise (S2N), is less well known. χ^2, IG, GR, RF, RFW and SU are available in the Weka data mining tool (Witten & Frank 2005). χ^2, IG, GR and SU utilize the method of Fayyad and Irani (Fayyad & Irani 1992) to discretize continuous attributes, and all four methods are bivariate, considering the relationship between each attribute and the class, excluding the other independent variables. S2N, also a bivariate method, was implemented by our research group for experimentation purposes since it is not available in Weka.

1) *Chi-Squared (χ^2)* is based on the χ^2-statistic and evaluates each feature independently with respect to the class labels. The larger the Chi-squared, the more relevant the feature is with respect to the class. Given the number of intervals (I), the number of classes (B), and the total number of instances (N), the Chi-squared value of a feature is computed as:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{B} \frac{[A_{ij} - \frac{R_i * B_j}{N}]^2}{\frac{R_i * B_j}{N}} \quad (1)$$

where R_i denotes the number of instances in the i^{th} interval, B_j the number of instances in the j^{th} class, and A_{ij} the number of instances in the i^{th} interval and j^{th} class.

2) *Information Gain (IG)* is a commonly used measure in the fields of information theory and machine learning. IG measures the number of bits of information gained about the class prediction when using a given feature to assist that prediction (Yang & Pedersen 1997). For each feature, a score is obtained based on how much more information about the class is gained when using that feature. The information gain of feature X is defined as:

$$IG(X) = H(Y) - H(Y|X) \quad (2)$$

where $H(Y)$ and $H(Y|X)$ are the entropy of Y and the conditional entropy of Y given X, respectively. The level of a feature's significance is thus determined by how great is the decrease in entropy of the class when considered with the corresponding feature individually.

3) *Gain Ratio (GR)* is a refinement to Information Gain. While IG favors features that have a large number of values, GR's approach is to maximize the feature's information gain while minimizing the number of its values. The gain ratio of X is thus defined as the information gain of X divided by its intrinsic value:

$$GR(X) = IG(X)/IV(X) \quad (3)$$

where $IV(X) = -\sum_{i=1}^{r} (|X_i|/N) log(|X_i|/N)$, from which $|X_i|$ is the number of instances where attribute X takes the value of X_i, r is the number of distinct values of X, and N is the total number of instances in the dataset.

4) *ReliefF (RF)* is an extension of the Relief algorithm introduced by Kira and Randell (Kira & Rendell 1992) and enhanced by Kononenko (Kononenko 1994). RF estimates the quality of a feature by finding one near miss ($M(B)$) for each different class and averages their contribution for updating estimates $W[X]$. The average is weighted with

the prior probability of each class, $P(B)$ (Kononenko 1994):

$$W[X] = \sum_{i=1}^{N} \left[\sum_{B \neq c(X)} \left[\begin{smallmatrix} P(B) \times d(X, I_0, M) \\ i \end{smallmatrix} \right] - \begin{smallmatrix} d(X, I_0, H) \\ i \end{smallmatrix} \right] \quad (4)$$

where $d(X, I_1, I_2)$ calculates the difference between the values of X for both instances I_1 and I_2.

5) *ReliefF-W (RFW)* is similar to ReliefF, except that in ReliefF-W, the "weight nearest neighbors by their distance" parameter is set to true.

6) *Symmetric Uncertainty (SU)* is a correlation measure between the features and the class (Witten & Frank 2005), and it is obtained by:

$$\begin{aligned} SU &= 2 \times \frac{H(X) + H(Y) - H(X, Y)}{H(X) + H(Y)} \\ &= 2 \times \frac{H(Y) - H(Y|X))}{H(X) + H(Y)} \end{aligned} \quad (5)$$

where $H(X)$ and $H(Y)$ are the entropies based on the probability associated with each feature and class value respectively and $H(X, Y)$, the joint probabilities of all combinations of values of X and Y.

7) *Signal to Noise (S2N)* is a simple univariate ranking technique which defines how well a feature discriminates two classes (Lakshmi & Mukherjee 2006). S2N is obtained for each feature using this formula:

$$S2N = \frac{\mu_+ - \mu_-}{\sigma_+ + \sigma_-} \quad (6)$$

where μ_+ and μ_- are the mean values for the feature from the positive class and negative class, respectively, and σ_+ and σ_- are the corresponding standard deviations.

Empirical Evaluation

Datasets

Seven datasets are considered in the empirical evaluation of the seven feature ranking techniques. The datasets represent different application domains, ranging from cancerous gene expression (Wang & Gotoh 2009) to drug activity, image recognition and Internet ad determination (Asuncion & Newman 2007). Table 1 lists all seven datasets and provides their characteristics in terms of the total number of attributes, number of instances, and percentage of positive instances. They are all binary class datasets with various class distribution levels.

To accomplish our goal of analyzing filters in the presence of class noise, noise is injected into the training datasets using two simulation parameters. These datasets are chosen because preliminary analysis showed near perfect classification. Ensuring that the datasets are relatively clean prior to noise injection is important because it is very undesirable to inject class noise into already noisy datasets.

Cross-validation and Runs

A 5-fold cross-validation procedure is used, where the set of N instances in the original dataset is randomly partitioned into 5 equal sets of size $N/5$. For each fold or partition which is used as test data, class noise is injected into the other 4 folds or partitions (i.e., the training data). Next, feature selection is performed on the noisy training data, and classification models are built with the selected features on the same noisy portions. Finally, the classification models are tested on the remaining 'clean' fold. This whole procedure is performed a total of 5 times using a different holdout 'clean' partition each time, and the results on each partition are combined to obtain a single performance metric. For each of the 5 times, the training dataset contains the noisy portions, representing 80% of the instances from the original dataset. Each classification model is tested on the clean portion, representing 20% of the instances from the original dataset, and each instance includes S_i top ranked features and the class. The overall cross-validation procedure is run four times. That is, the random partioning of the original datasets into 5 equal folds is repeated three more times, allowing for four distinct observations.

Dataset	ID	#Attributes	#Instances	%Positive
Lung Cancer	LC	12534	181	17.1
Ovarian Cancer	OC	15155	253	36.0
Liver Cancer	VC	122	166	47.0
Internet Ad	IA	1559	3279	14.0
Musk	MK	167	6598	15.4
Satimage-4	S4	65	5620	9.9
Optdigits-8	O8	37	6435	9.7

Table 1: Dataset Characteristics

Noise Injection Mechanism

Class noise is injected into the datasets, but only in the four partitions intended for training. For the noise injection mechanism, the same procedure as reported by (Van Hulse & Khoshgoftaar 2009) is used. That is, the levels of class noise are regulated by two noise parameters. The first parameter, denoted α ($\alpha = 10\%, 20\%, 30\%, 40\%, 50\%$), is used to determine the overall class noise level (NL) in the data. Precisely, α is the noise level relative to the number of instances belonging to the positive class, i.e., the number of examples to be injected with noise is $2 \times \alpha \times \mid P \mid$, where $\mid P \mid$ is the number of examples in the smaller class (often referred to as the positive class). This ensures that the positive class is not drastically impacted by the level of corruption, especially if the data is highly imbalanced. The second parameter, denoted β ($\beta = 0\%, 25\%, 50\%, 75\%, 100\%$), represents the percentage of class noise injected in the positive instances and is referred to as noise distribution (ND). In other words, if there are 125 positive class examples in the training dataset and $\alpha = 20\%$ and $\beta = 75\%$, then 50 examples will be injected with noise, and 75% of those (38) will be from the positive class. These parameters serve to ensure systematic control of the training data corruption. Due to space constraints, more details on the noise injection scheme are not included. For those details, readers are referred to (Van Hulse & Khoshgoftaar 2009).

Feature Selection

Once the features are ranked according to their relevance to the class, a subset consisting of the most relevant ones is selected. For each dataset, the specified number of features that is retained is denoted by S_i, where i represents one of the seven datasets. Table 2 shows the value of this parameter for each dataset. These values were selected based on preliminary experimentation on each dataset and were deemed reasonable for the corresponding dataset. The intent is not to investigate different number of attributes, nor is it to analyze the 'optimal' choice for the number of attributes. Our objective here is to select values that are reasonable and perform generally well across a wide range of experimental conditions to allow us to effectively compare the different feature ranking techniques.

Dataset	S_i	Percentage
Lung Cancer	31	0.25
Ovarian Cancer	38	0.25
Liver Cancer	12	10
Internet Advertisements	78	5
Musk	34	20
Satimage-4	13	35
Optdigits-8	19	30

Table 2: Thresholds

Classification Algorithms and Performance Measure

After selecting the S_i relevant features for each dataset, three different machine learning algorithms are used for the classification models: Naïve Bayes (NB), k-Nearest Neighbors (kNN) with k set to 5, and Support Vector Machines (SVM). These algorithms, commonly used in data mining, are readily available in the Weka data mining tool (Witten & Frank 2005), which is used in this study. The classification performance of each algorithm is evaluated by the area under the receiver operating characteristic curve (AUC) metric, which is chosen because of its invariance to a priori class probability distributions and its statistical consistency (Jin *et al.* 2003). The classification results are then used to compare the ranking techniques in terms of their robustness to class noise.

Results

Tables 3 – 5 show the classification performances in terms of AUC on a particular classifier for all the filters and for all noise injection schemes. The first two columns of each figure represent the noise injection schemes (combinations of noise levels and noise distributions). Each entry in the remaining columns represents the result in terms of the average AUC over all 7 datasets ($1 \leq d \leq 7$), for a particular filter, i, and a particular noise injection scheme, n, $AUC_{i,n} = \frac{1}{7}\sum_{d=1}^{7} AUC_{i,n}^{d}$. For instance, the value .955 shown under the column header χ^2 in Table 4 indicates the average classification performance obtained from 5NN after χ^2 has been applied to the training data which has been injected with noise corresponding to a 10% noise level and a 0% noise distribution. Likewise, the average classification performance of S2N with 5NN for the noise injection scheme consisting of a 50% noise level and a 75% noise distribution is .690.

NL	ND	χ2	GR	IG	RF	RFW	SU	S2N
10%	0%	.938	.930	.940	.939	.937	.938	.933
	25%	.939	.930	.942	.939	.940	.940	.936
	50%	.942	.930	.943	.942	.938	.943	.938
	75%	.942	.930	.945	.940	.936	.943	.939
	100%	.941	.931	.944	.940	.934	.945	.937
20%	0%	.935	.925	.938	.938	.931	.934	.927
	25%	.937	.924	.941	.939	.925	.937	.931
	50%	.937	.923	.940	.935	.925	.937	.932
	75%	.937	.919	.941	.932	.924	.938	.934
	100%	.937	.919	.940	.935	.928	.940	.936
30%	0%	.928	.919	.931	.935	.929	.927	.920
	25%	.931	.917	.932	.933	.907	.929	.922
	50%	.928	.890	.931	.924	.901	.926	.922
	75%	.924	.887	.926	.922	.907	.925	.923
	100%	.925	.881	.932	.932	.920	.928	.925
40%	0%	.918	.908	.921	.923	.919	.917	.911
	25%	.906	.867	.908	.911	.884	.904	.907
	50%	.898	.847	.898	.901	.863	.886	.901
	75%	.884	.846	.888	.902	.871	.875	.892
	100%	.865	.852	.895	.914	.909	.874	.898
50%	0%	.895	.883	.897	.915	.906	.894	.889
	25%	.878	.831	.877	.865	.832	.864	.873
	50%	.837	.803	.836	.854	.796	.828	.846
	75%	.799	.782	.799	.845	.776	.797	.806

Table 3: Filters' Performances for All Noise Injection Schemes on NB

At a high noise level (40% or 50%), the classification performance generally decreases as the noise distribution increases for most filters on all three classifiers. For a fixed noise distribution, the performance decreases as the noise level increases. The results also show that the filters perform better on 5NN at lower noise levels (10% and 20%) while at higher noise levels (30%, 40%, and 50%), NB exhibits better performance. In other words, NB is generally more robust to class noise than 5NN, regardless of the filter. With SVM, the classification performance generally degrades for all noise injection schemes in comparison to NB and 5NN. For more than 85% of the noise injection schemes, GR shows the lowest classification performance when either NB or 5NN is the classifier. Conversely, for about 90% of the noise injection schemes, the filter that has the best classification performance with NB and 5NN is either IG or RF. With SVM as the classifier, GR and RFW are the least effective while S2N proves to be the most effective.

Robustness

The robustness of each filter to class noise is measured by obtaining the sum of the squared differences between 1 (for a perfect classification) and the corresponding classification performance. This measure is equivalent to the sum of squared errors (SSE) and is obtained by: $SSE_i^d = \sum_{n=1}^{24}(1 - AUC_{i,n}^{d})^2$. The data is provided on a per-dataset

NL	ND	χ2	GR	IG	RF	RFW	SU	S2N
10%	0%	.955	.941	.955	.959	.956	.951	.942
	25%	.956	.939	.956	.957	.955	.954	.942
	50%	.956	.938	.956	.957	.954	.953	.940
	75%	.957	.940	.958	.955	.954	.955	.941
	100%	.956	.941	.958	.957	.955	.956	.942
20%	0%	.949	.937	.950	.955	.950	.945	.939
	25%	.945	.930	.945	.947	.940	.943	.932
	50%	.941	.924	.942	.939	.933	.939	.927
	75%	.943	.919	.944	.939	.933	.940	.923
	100%	.944	.916	.946	.945	.942	.943	.922
30%	0%	.938	.929	.941	.939	.936	.938	.928
	25%	.928	.914	.929	.931	.919	.929	.915
	50%	.915	.872	.914	.917	.904	.913	.901
	75%	.907	.861	.907	.904	.898	.903	.883
	100%	.912	.854	.918	.918	.917	.910	.889
40%	0%	.917	.906	.919	.917	.916	.916	.909
	25%	.886	.851	.887	.895	.886	.882	.881
	50%	.866	.810	.864	.874	.852	.856	.851
	75%	.827	.788	.833	.843	.832	.818	.818
	100%	.804	.752	.816	.855	.853	.791	.786
50%	0%	.881	.867	.882	.896	.883	.878	.873
	25%	.845	.801	.846	.845	.826	.832	.843
	50%	.788	.753	.793	.810	.782	.777	.776
	75%	.706	.687	.714	.748	.724	.705	.690

Table 4: Filters' Performances for All Noise Injection Schemes on 5NN

NL	ND	χ2	GR	IG	RF	RFW	SU	S2N
10%	0%	.901	.901	.902	.899	.896	.900	.905
	25%	.903	.893	.899	.901	.901	.902	.908
	50%	.897	.894	.901	.898	.896	.898	.908
	75%	.897	.892	.899	.897	.899	.900	.902
	100%	.897	.887	.903	.889	.890	.899	.893
20%	0%	.895	.891	.896	.893	.889	.895	.901
	25%	.897	.890	.898	.889	.884	.896	.902
	50%	.892	.880	.893	.884	.876	.894	.895
	75%	.882	.871	.887	.875	.873	.881	.892
	100%	.875	.846	.878	.872	.862	.866	.875
30%	0%	.887	.884	.887	.884	.877	.892	.893
	25%	.885	.870	.889	.875	.862	.885	.886
	50%	.874	.842	.873	.847	.837	.869	.865
	75%	.819	.787	.825	.815	.796	.819	.821
	100%	.788	.769	.795	.774	.768	.779	.784
40%	0%	.877	.871	.877	.861	.865	.878	.879
	25%	.859	.829	.862	.848	.831	.864	.847
	50%	.793	.765	.796	.779	.765	.789	.796
	75%	.699	.711	.696	.708	.704	.700	.711
	100%	.710	.706	.735	.721	.705	.721	.750
50%	0%	.852	.848	.854	.845	.848	.855	.872
	25%	.818	.771	.816	.789	.766	.801	.806
	50%	.702	.698	.701	.674	.664	.693	.710
	75%	.598	.596	.596	.604	.600	.600	.604

Table 5: Filters' Performances for All Noise Injection Schemes on SVM

basis and over all seven datasets in Tables 6 to 8, with one table for each learner. The first column lists the filters ordered from the most to the least robust. Each filter's overall robustness is shown in the second column. The remaining columns show the robustness for each filter according to each dataset. Contemplating the results on all three classifiers, the most unstable filter is GR given its highest SSE values. Comparing the robustness of S2N with the standard filters, the results show that S2N outperforms χ^2, SU, RFW and GR for two of the three classifiers (NB and SVM). By contrast, S2N's performance on 5NN is rather poor.

Clearly, the effectiveness of a filter depends on the data as well as the classifier used in the experiments. Nonetheless, when dealing with poor quality data, GR, though a widely used filter, is not recommended. S2N, a rarely used filter, shows some potential in terms of its robustness to class noise.

Filter	All	LC	OC	VC	IA	MK	S4	O8
RF	**.175**	.018	**.051**	.893	.250	**.385**	.273	**.036**
IG	.200	.023	.222	.935	.078	.747	**.241**	.049
S2N	.212	**.015**	.111	**.752**	.206	.739	.324	.056
χ2	.215	.047	.237	.933	**.076**	.828	.287	.050
SU	.225	.046	.289	.940	.076	.846	.264	.055
RFW	.278	.434	.130	1.017	.278	.487	.268	.036
GR	.331	.072	.503	.946	.456	.854	.266	.065

Table 6: SSE on NB per Dataset

Filter	All	LC	OC	VC	IA	MK	S4	O8
RF	**.270**	**.020**	**.240**	1.357	.261	.335	.655	**.017**
RFW	.325	.151	.261	1.554	.295	.362	**.618**	.019
IG	.318	.050	.373	1.576	**.195**	**.329**	.697	.052
χ2	.334	.072	.413	1.588	.199	.338	.701	.053
SU	.360	.099	.476	1.597	.198	.365	.713	.070
S2N	.401	.065	.352	**1.274**	.714	.411	.786	.107
GR	.506	.263	.746	1.609	.637	.383	.709	.091

Table 7: SSE on 5NN per Dataset

Filter	All	LC	OC	VC	IA	MK	S4	O8
S2N	**.746**	**.175**	**.203**	**.873**	1.164	1.309	**3.320**	.455
IG	.770	.176	.257	1.016	1.165	**.891**	3.930	.520
χ2	.791	.176	.281	1.035	1.149	.991	3.979	.470
SU	.800	.221	.274	1.024	1.126	1.154	3.753	**.451**
RF	.847	.276	.279	1.110	1.174	1.024	4.026	.601
GR	.891	.312	.319	1.027	1.256	1.288	3.809	.506
RFW	.913	.453	.254	1.123	**1.039**	1.179	4.054	.635

Table 8: SSE on SVM per Dataset

Conclusion

Studies on feature ranking techniques have traditionally used classification performances from models built with a subset of the original features to assess the strengths and weaknesses of the techniques. However, there has been no comparative study of feature ranking techniques that has taken into consideration the impact of class noise on the performance of the filters. In this study, a method for comparing the filters' robustness is introduced. It involves obtaining

classification performances through 5-fold cross validation, whereby feature selection is performed on the portions of the data injected with class noise and classification models are tested with the remaining clean portion. For the empirical evaluation, seven feature ranking techniques are used. Six of these techniques are commonly used and are referred to as standard filters: Chi-squared, Information Gain, Gain Ratio, two versions of ReliefF and Symmetric Uncertainty. The remaining filter is Signal-to-noise, a ranking technique that is rarely used. The classification performances of three different learners (NB, 5NN and SVM) in terms of AUC are used to compare the filters' robustness to class noise.

Using seven binary classification datasets representing different application domains and different class distribution levels, each filter's robustness against class noise is measured on all three classifiers in terms of SSE. The empirical results show that S2N, although rarely used, can outperform some of the widely used filters. On all three classifiers, S2N demonstrates more stability than GR.

Given the results of these experiments, GR proves to be the least effective among the seven filters. On the other hand, the IG filter stands out in terms of its robustness to class noise, given its ranking among the top three for all classifiers. Considering the performance of the filters on individual classifiers, RF performs best with both NB and 5NN, while S2N outperforms with SVM. S2N, though not as widely used in literature as the others, is preferred over most when the learning algorithm is either NB or SVM. Thus, both the filters and the learning algorithms are affected by class noise. In the presence of class noise, the most impacted filter and learning algorithm are GR and SVM, respectively. The best overall three filters are: RF, IG and S2N.

While this paper only considers class noise in the empirical evaluation, future work could assess the impact of attribute noise on the feature ranking techniques' performances. Other noise injection schemes could also be considered.

References

Asuncion, A., and Newman, D. 2007. UCI machine learning repository [http://www.ics.uci.edu/∼mlearn/MLRepository.html]. University of California, Irvine, School of Information and Computer Sciences.

Fayyad, U. M., and Irani, K. B. 1992. On the handling of continuous-valued attributes in decision tree generation. *Machine Learning* 8:87–102.

Gilad-Bachrach, R.; Navot, A.; and Tishby, N. 2004. Margin based feature selection - theory and algorithms. In *International Conference on Machine Learning (ICML)*, 43–50. ACM Press.

Guyon, I., and Elisseeff, A. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3:1157–1182.

Hall, M. A., and Holmes, G. 2003. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering* 15(6):1437–1447.

Jin, C. L.; Ling, C. X.; Huang, J.; and Zhang, H. 2003. Auc: a statistically consistent and more discriminating measure than accuracy. In *Proceedings of 18th International Conference on Artificial Intelligence*, 329–341.

Kira, K., and Rendell, L. A. 1992. A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning*, 249–256. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Kononenko, I. 1994. Estimating attributes: Analysis and extensions of RELIEF. In *European Conference on Machine Learning*, 171–182. Springer Verlag.

Lakshmi, K., and Mukherjee, D. S. 2006. An improved feature selection using maximized signal to noise ratio technique for TC. In *Proceedings of the Third International Conference on Information Technology: New Generations*, 541–546. Washington, DC, USA: IEEE Computer Society.

Liu, H.; Li, J.; and Wong, L. 2002. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics* 13:51–60.

Méndez, J. R.; Fdez-Riverola, F.; Díaz, F.; Iglesias, E. L.; and Corchado, J. M. 2006. A comparative performance study of feature selection methods for the anti-spam filtering domain. In *Industrial Conference on Data Mining*, 106–120.

Ruiz, R.; Aguilar-Ruiz, J. S.; Santos, J. C. R.; and Díaz-Díaz, N. 2005. Analysis of feature rankings for classification. In *IDA*, 362–372.

Saeys, Y.; Inza, I. n.; and Larrañaga, P. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23(19):2507–2517.

Van Hulse, J., and Khoshgoftaar, T. 2009. Knowledge discovery from imbalanced and noisy data. *Data and Knowledge Engineering* 68(12):1513–1542.

Wang, X., and Gotoh, O. 2009. Accurate molecular classification of cancer using simple rules. *BMC Medical Genomics* 2(1):64+.

Wang, H.; Khoshgoftaar, T.; and Gao, K. 2010. Ensemble feature selection technique for software quality classification. In *Proceedings of the 22nd International Conference on Software Engineering and Knowledge Engineering*, 215–220.

Witten, I. H., and Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2 edition.

Xiong, H.; Pandey, G.; Steinbach, M.; and Kumar, V. 2006. Enhancing data analysis with noise removal. *IEEE Transactions on Knowledge and Data Engineering* 18(3):304–319.

Yang, Y., and Pedersen, J. O. 1997. A comparative study on feature selection in text categorization. In Fisher, D. H., ed., *Proceedings of ICML-97, 14th International Conference on Machine Learning*, 412–420. Nashville, US: Morgan Kaufmann Publishers, San Francisco, US.