Comparison of Redundancy and Relevance Measures for Feature Selection in Tissue Classification of CT images*

Benjamin Auffarth[1]**, Maite López[2], and Jesús Cerquides[2]

1 Institute for Bioengineering of Catalonia
C/Baldiri Reixac 4-6 (torre I), 08028 BCN, Spain
2 Volume Visualization and Artificial Intelligence research group,
Departament de Matemàtica Aplicada i Anàlisi (MAIA), Universitat de Barcelona,
C/Gran Via, 585, 08007 Barcelona, Spain

bauffarth@el.ub.es,{maite,jcerquide}@maia.ub.es

Abstract. In this paper we report on a study on feature selection within the minimum–redundancy maximum–relevance framework. Features are ranked by their correlations to the target vector. These relevance scores are then integrated with correlations between features in order to obtain a set of relevant and least–redundant features. Applied measures of correlation or distributional similarity for redunancy and relevance include Kolmogorov–Smirnov (KS) test, Spearman correlations, Jensen–Shannon divergence, and the sign–test. We introduce a metric called "value difference metric" (VDM) and present a simple measure, which we call "fit criterion" (FC). We draw conclusions about the usefulness of different measures. While KS–test and sign–test provided useful information, Spearman correlations are not fit for comparison of data of different measurement intervals. VDM was very good in our experiments as both redundancy and relevance measure. Jensen–Shannon and the sign–test are good redundancy measure alternatives and FC is a good relevance measure alternative.

Key words: feature selection; relevance and redundancy; distributional similarity; divergence measure

1 Introduction

In biomedical image processing, it is difficult to classify organ tissues using shape or gray level information, because image intensities overlap considerably for soft tissue. Hence, features used for processing often go beyond intensity and include

* This research was supported by the Spanish MEC Project "3D Reconstruction, classification and visualization of temporal sequences of bioimplant Micro-CT images" (MAT-2005-07244-C03-03).

** Corresponding author.

something what can be very generally referred to as texture (see [1]). The use of an adequate feature set is a requirement to achieve good classification results.

Feature selection generally means considering subsets of features and eventually choosing the best of these subsets. The "goodness" of feature subsets can be estimated by filters, such as statistical or information theoretic measures, or by a performance score of a classifier (*wrappers*). Currently many approaches to feature selection in bioinformatics are either based on rank filters (univariate filter paradigm) and thereby do not take into account relationships between features, or are wrapper approaches which require high computational costs.

Multivariate filter-based feature selection has enjoyed increased popularity recently [2]. The approach is generally low on computational costs. Filter–based techniques provide a clear picture of why a certain feature subset is chosen through the use of scoring methods in which inherent characteristics of the selected set of variables is optimized. In comparison wrapper-based approaches treat selection as a black-box and optimize the prediction ability according to a chosen classifier.

In feature selection, it is important to choose features that are relevant for prediction, but at the same time it is important to have a set of features which is not redundant in order to increase robustness. [3,4,5,6,7,8,9,10] have elaborated on the concepts of redundancy and relevance for feature selection. [11,4,9] presented feature selection in a framework they call min-redundancy max-relevance (here short mRmR) that integrates relevance and redundancy information of each variable into a single scoring mechanism.

Our data consist of slices in a 3-D volume taken from CT of bones, into which a tracing material was introduced[1]. Fig. 1 shows the alien biomaterial marked in white on the right and organic bone material (referred to henceforth as non-biomaterial in order to distinguish it from the introduced biomaterial). For the classification task, the introduced biomaterial is the target class and relatively small as compared to the non-target class. The volume centers around the introduced material and hence, percentage of biomaterial is greatest in the centers (around 10 percent), becoming less towards the exteriors.

In this article we present an experimental evaluation of several filters for computing redundancy and relevance. In section 2 we will introduce the concepts of redundancy and relevance and compare several measures. We put emphasis on non–parametric filters that are low on computational costs, using very simple density estimation. Section 3 describes experimental methodology and the results are presented in section 4. In section 5 we then discuss and draw some conclusions.

2 Feature Selection with Relevance and Redundancy

In feature selection the aim is to choose a subset of features in order to improve performance and efficiency with respect to a task (in our case classification) and

[1] Samples from the data set are available on the homepage of one of the authors: `http://www.maia.ub.es/~maite/out-slice-250-299.arff`.

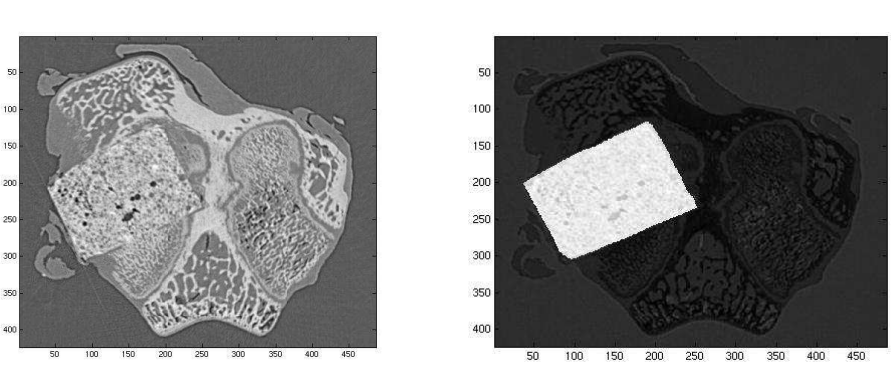

Fig. 1. A central slice in the 3-D volume (left) and its labels (right) marked white

to reduce noise. Information loss in the reduction of the feature space should be kept as small as possible so the resulting space can provide enough information for classification.

Relevance measures the "goodness" of the projection from individual attributes to labels. Redundancy measures how similar features are (or inversely, how much adding a feature to a given set of features contributes to prediction[2]). As such, both redundancy and relevance measures fall into the class of measures for *statistical dependence*, *distributional similarity*, or *divergence measure*.[3]

We will now outline several relevance and redundancy criteria based on mutual information, statistical tests, probability distributions, and correlation coefficients. Since we do not know, the true distribution of the data, we prefer non-parametric and model-free metrics. Non-parametric tests have less power (i. e. the probability that they reject the null hypothesis is smaller) but should be preferred when distributions could be non–gaussian. Furthermore non–parametric filters are generally more robust to outliers than parametric tests.

Within the context of feature selection, we will write targets as $Y \in C^N, |C| = d, C = \{c1, ..., c_d\}$ and denote feature i as X_i, with elements x_i^k.

2.1 Relevance Criteria

Relevance is distributional similarity between a continous feature vector and a target vector. In this article we consider the case of two classes (binary classification). Relevance criteria determine how well a variable discriminates between

[2] Even though, redundancies can be n-ary relations of features, henceforth we will take redundancies to mean binary relations, i. e. between only two features, which is how it was used in [11,4,9].

[3] Shortly, while similarity and divergence are two different concepts, in this context, divergence or distance is taken to mean dissimilarity.

the classes. They are a measure between a feature and the class, i. e.

$$Rel(X, Y) \equiv \text{ how useful is } X \text{ for predicting } Y. \tag{1}$$

The relevance criteria that we discuss and use later in experiments are:

- Symmetric Uncertainty (SU)
- Spearman rank correlation coefficient (CC)
- Value Difference Metric (VDM)
- Fit Criterion (FC)

Of these, symmetric uncertainty was used before as a relevance criterion [5,7]. Symmetric uncertainty is symmetric and scaled mutual information [12]. Mutual information was used by [4,9] and by [3]. [13] used normalized mutual information for gene selection.

As for Spearman correlations, we did not find a prior publication that refers to it as a relevance criterion, but we thought it might be better to use a non-parametric measure instead of relying on linear correlations (Pearson product–moment correlations), which have been used before as relevance measure [14,8]. We did not use Pearson correlations because of their sensibility to extreme values, their focus on strictly linear relationships, and the assumption of gaussianity. For non-gaussian data rank–correlations should be preferred over Pearson correlations (see [15] on rank correlations and [16] as one of many recommendations to use Spearman correlations instead).

We show how a measure of probability difference, similar to one presented before as the "value difference metric" [17], can be adapted as a relevance criterion. We propose a new measure, which we call "fit criterion" which measures relevance similar to the z-score.

Value Difference Metric We will refer to $\mathrm{p}(X)$ as the probability function of variable X, $\mathrm{p}([X|Y = c_i])$ as the probability function of X with target $Y = c_i$, and $\mathrm{p}(X = x)$ as the probability density of X at x.

We define a simple, continuous, monotonic function that measures overlap between two variables X_1 and X_2:

$$\left(\int |\mathrm{p}(X_1 = x) - \mathrm{p}(X_2 = x)|^q \, dx \right)^{1/q}, \text{ where } q \text{ is a parameter} \tag{2}$$

We chose $q = 1$, which has been used similarly by [17,18,19] under the name *value difference metric* as distance measure.

Given that the probabilities that X is equal to a given x for all possible values of x is 1, $\int \mathrm{p}(x)\, dx = 1$, total divergence would give the sum

$$\int \mathrm{p}(X_1 = x)\, dx + \int \mathrm{p}(X_2 = x)\, dx = 2 \tag{3}$$

In order to have a range between 0 and 1 we divided by 2. This gives a very intuitive, vertical distance between the probability mass functions.

$$\mathrm{VDM}(X_1, X_2) = \frac{1}{2} \int |\mathrm{p}(X_1 = x) - \mathrm{p}(X_2 = x)| \; dx \tag{4}$$

Our VDM relevance measure is based on the idea that conditional distributions of variables $\big\{p([X_i|Y = c_j])\big|j = 1, \ldots, d\big\}$ should be distinct from each other. We define VDM relevance (to which we will refer to short as VDM) of a feature X and labels Y with two classes c_1 and c_2 as:

$$\mathrm{VDM}(X, Y) = \frac{1}{2} \int |\mathrm{p}(X = x|c_1) - \mathrm{p}(X = x|c_2)| \; dx \tag{5}$$

Fit Criterion For a given point x a criterion of fit to one distribution X_1 could be defined as the points distance to the center of the distribution $\overline{X}_1$ in terms of the variance of the distribution var_{X_1}.

$$\frac{\big|x - \overline{X}_1\big|}{\mathrm{var}_{X_1}} \tag{6}$$

where $\overline{X}$ is a center of the distribution (as given e. g. by the mean or median[4]).) and var denotes some measure of statistical dispersion, (e. g. the mean absolute deviation from the mean)

A decision criterion for whether a point x belongs to distribution X_1 or to distribution X_2 could be this:

$$\mathrm{FCP}(x, X_1, X_2) = \begin{cases} 1 \text{ if } \frac{|x-\overline{X}_1|}{\mathrm{var}_{X_1}} < \frac{|x-\overline{X}_2|}{\mathrm{var}_{X_2}} \\ 2 \text{ if } \frac{|x-\overline{X}_1|}{\mathrm{var}_{X_1}} > \frac{|x-\overline{X}_2|}{\mathrm{var}_{X_2}} \end{cases} \tag{7}$$

In the case that both distances were equal we chose arbitrarily.

We refer to FCP as the *fit criterion for a given point*.

More general for k distributions and a feature, this can be expressed as

$$\mathrm{FCP}(x, X) = \mathrm{arg}_{i=1\ldots,k} \min \frac{\big|x - \overline{X}_i\big|}{\mathrm{var}_{X_i}} \tag{8}$$

We now show the derivation of the decision boundary $\dot{x}$ that results from FCP given again two distributions X_1 and X_2. Our decision boundary $\dot{x}$ is at equal distance to both μ_{X_1} in terms of σ_{X_1} and μ_{X_2} in terms of σ_{X_2}.

$$\frac{|\mu_{X_1} - \dot{x}|}{\sigma_{X_1}} = \frac{|\mu_{X_2} - \dot{x}|}{\sigma_{X_2}} \tag{9}$$

[4] The choice between mean and median should depend on characteristics of the data and the task. However, as the classical center of gravity the mean is preferable.

We also know that $\dot{x}$ is between μ_{X_1} and μ_{X_2}. We assume $\mu_{X_1} \leq \mu_{X_2}$ and therefore $\mu_{X_1} \leq \dot{x} \leq \mu_{X_2}$ and resolve

$$\dot{x} = \frac{\mu_{X_1}\,\sigma_{X_2} + \sigma_{X_1}\,\mu_{X_2}}{\sigma_{X_2} + \sigma_{X_1}} \text{ (if } \mu_{X_1} \leq \mu_{X_2}) \tag{10}$$

Such decision boundaries ignore many of the characteristics of the distributions, but are unbiased between different distributions, because they do not take into account prior class-probabilities. Note that the above expression loses meaning with long tails and with n-modal distributions ($n > 1$).

For the decision, whether x from distribution X_1 belongs to class c_1 or c_2 we write $\mathrm{FCP}(x, [X_1|Y = c_1], [X|Y = c_2])$.

For calculating relevance based on the FCP, we proceed with the conditional distributions $p([X_i|Y = c_1])$, X_i, where corresponding targets are equal to c_1, and for each point $x \in X_i$ we compute the FCP, i. e. the class which point x should belong to according to equation 8. This is then matched with the target labels and the percentage of correct classification by equation 8. We use this as a relevance criterion and call it "fit criterion" (short "FC"). Given data $[X|Y = c_i]$ of features $X = \left\{x_i^j \middle| i = 1 \ldots m, j = 1 \ldots N\right\} \subset \mathbb{R}^{Nm}$, where N is the number of points and m the number of features, and matching class labels $Y = \left\{y^j \middle| j = 1 \ldots N\right\} \in C^N$, we define the relevance fit criterion for binary class labels in Y and some feature X_k as:

$$\mathrm{FC}(X_k, Y) = \frac{1}{n}\sum_{i=1}^{N} 1_{\mathrm{FCP}(x_k^i, [X_k|y^i=c_1], [X_k|y^i=c_2])=y^i}, \tag{11}$$

where 1 is an indicator function returning 1 (correct) or 0 (incorrect) depending on the correctness of the prediction by FCP. This relevance criterion takes the average accuracy of the separation by the σ–normalized distance from centers of distribution X_k given label c_1 and given label c_2, respectively.

2.2 Redundancy Criteria

Redundancy criteria measure similarity between the distribution of attributes and the distribution of labels[5].

Formally the redundancy between features X_1 and X_2 given class targets $Y \in C^N = \{c_1, \ldots, d\}^N$ can be written as

$$\mathrm{Red}(X_1, X_2, Y) = \frac{1}{d}\sum_{i=1}^{d} \Delta\left([X_1|Y = c_i], [X_2|Y = c_i]\right), \tag{12}$$

where $[X_1|Y = c_i]$ denotes the distribution of feature 1, given class i (i. e. $\{X_1^l|\forall l, Y^l = c_i\}$, and Δ one of the distributional similarity measures that will

[5] As such there exists abundant literature on goodness of fit however we did not find comparisons within the context of feature selection for pattern recognition.

follow in this subsubsection. There could be more advantageous ways to combine the conditional metrics than the arithmetic mean as in equation 12, but we chose consciously a conservative one.

Relevance and redundancy measures are tests for the goodness-of-fit and as such, we could use similar or even the same functions for measuring redundancy and relevance. Given a relevance measure $Rel()$, features X_1 and X_2, and targets $Y \in C^N$, we can define

$$\mathrm{Red}(X_1, X_2, Y) = \frac{1}{d} \sum_{i=1}^{d} \left(\mathrm{Rel}([X_1|Y = c_i], [X_2|Y = c_i])\right). \tag{13}$$

We used these redundancy criteria:

- Kolmogorov-Smirnov test on class-conditional distributions (RKSC)
- Kolmogorov-Smirnov test ignoring classes (RKSD)
- Redundancy VDM (RVDM)
- Redundancy Fit Criterion (RFC)
- Spearman rank correlation coefficients (RCC)
- Jensen-Shannon Divergence (RJS)
- Sign–test (RST)

Redundancy can be measured taking into account classes or without respect to a given class. For purpose of comparison, we include two redundancy criteria that differ only in whether or not they use class information, RKSC and RKSD. We compute all redundancy measures on the class conditional distributions except for RKSD. Recently, Zhang et al. found that taking class–specific correlations they obtained better results.

The Jensen-Shannon divergence is a symmetric and scaled version of the Kullback-Leibler divergence (sometimes: information divergence, information gain, relative entropy, which is an information theoretic measure of the difference between two probability distributions P and Q [20].

We will describe the redundancy VDM and the redundancy fit criterion in the following.

Redundancy Fit Criterion Equation 11 gives the goodness of fit with respect to two classes, c_1 and c_2, averaged over all points of a feature X_k. The binary sequence behind the sum represents correct class attributions (hits, 1) and incorrect class attributions (misses, 0) for each point of a feature X_k. Let us write the indicator function (and binary vector) corresponding to feature X_k as $\mathrm{hits}_{\mathrm{X_k}} \in \{1, 0\}^N$, where N are the number of points of X_k. We define hits as:

$$\mathrm{hits} \begin{cases} 1 \text{ if } \mathrm{FCP}(x_k^i, \left[X_k|y^i = c_1\right], \left[X_k|y^i = c_2\right]) = y^i \\ 0 \qquad\qquad\qquad\qquad \text{otherwise} \end{cases} \tag{14}$$

A very simple similarity measure between two features X_1 and X_2 given their binary sequences hits_{X_1} and hits_{X_2} could be the normalized sum of hits

combined by binary operators:

$$\mathrm{RFC}_{X_1,X_2} = \frac{\sum (\mathrm{hits}_{X_1} \wedge \mathrm{hits}_{X_2}) \vee (\neg\mathrm{hits}_{X_1} \wedge \neg\mathrm{hits}_{X_2})}{N} \tag{15}$$

This formula quantifies the percentage of identically classified points. We will refer to this measure as the redundancy fit criterion“ (short “RFC“).

3 Experiments

We benchmarked the feature selection quality resulting from redundancy and relevance information combined by different selection schemes. Additionally we selected features based on unitary filters, i. e. based on either relevance or redundancy. We benchmarked first each relevance and redundancy criterion on its own by unitary filters, then all 28 combinations of mentioned relevance and redundancy measures with different selection schemes and random selection. We selected feature sets of different sizes ($S = [4, 8, 12, 16, 20, 30, 45, 60, 80, 100]$[6]).

We compared five basic feature selection schemes. In the simplest selection scheme, at each iteration we take the most relevant feature and discard all features for which redundancy with the newly chosen features exceeds a threshold. We iterate over these two steps until no features are left. This scheme was presented by [5] and we refer to it henceforth as “Greedy“. Varying the redundancy thresholds we obtain a different number of features. As for the second selection scheme we order features by either $\frac{\mathrm{rel}}{\mathrm{red}}$ or $\mathrm{rel} - \mathrm{red}$ and choosing the first s. This schemes, presented by [9,4] were called minimum redundancy maximum relevance quotient (mRmRQ) and minimum redundancy maximum relevance difference (mRmRD), respectively. In [21] we presented a selection scheme based on an attractor network, which was thought to be capable of integrating more complex redundancy interactions between features (henceforth called Hopfield) and was comparable in performance to the mRmR framework. This is our third selection scheme. As our last selection scheme we rely on unitary filters which means either only relevance or only redundancy was taken into account. For the relevance case, the s most relevant features were used, and for the redundancy case, starting from the complete set of features, at each step the most redundant feature is removed until the desired numbers of features s are left. At last, we also compared a baseline of random selection.

We applied three classifiers for benchmarking. These were Naïve Bayes, GentleBoost, and a linear Support Vector Machine. As for Naïve Bayes we relied on our own implementation for multi-valued attributes using 100 bins. For GentleBoost we used 50 iterations. For SVM classification, we used libsvm [22]. Features were z-normalized and the cost function was made to compensate for unequal class priors, i. e. the weight of the less frequent class was set to $\max\left(\frac{\sharp(Y=c_2)}{\sharp(Y=c_1)}, \frac{\sharp(Y=c_1)}{\sharp(Y=c_2)}\right)$. We set the SVM complexity parameter C to 1 which seemed to be a good choice and in the right order of magnitude.

[6] This choice expresses an emphasis on feature sets of sizes ≤ 30 because that was were they were the greatest differences between the different methods.

At each number of features – in order to have many validations at acceptable speed – we made 10 random samplings of size $n/10$ and for each sampling we did 5-fold cross-validation. As for random feature selection, we did 10 random samplings of the data of size $n/10$ and tested 10 random selections of features in 5-fold cross-validation.

The complete feature set consisted of 127 features. We included 10 features from the Laplacian Pyramid [23], 100 Gabor features [24] in 10 orientations and 10 scales, 9 features from luminance contrast [25], 7 features from texture contrast [26], and intensity.We added 50 useless variables (probes) which good feature selection methods should eliminate. 49 of these probes were random variables. 25 of those standard normal distributed, 24 uniformly distributed in the interval $(0, 1)$. The last probe was a variable of zeros.

The experiments and comparisons following in this section are therefore based on a set of 177 features and their respective relevance measures and mutual redundancies. Details on the methods can be found in [27].

4 Results

Within the scope of this article we focus on these questions:

1. What are the best measures of relevance and redundancy (RR)?
   (a) What is the best redundancy and relevance (RR) combination?
   (b) What is the best redundancy measure?
   (c) What is the best relevance measure?
2. Do class-conditional distributions give better redundancy estimations?

Question 1 concerns comparisons of relevance and redundancy measures. In particular this concerns comparisons of combinations of redundancy and relevance measures, and of redundancy measures and relevance measures, respectively, among themselves.

In subsection 2.2 we proposed to calculate redundancy criteria based on class-conditional distributions. As for question 2, we want to resolve whether this made sense, looking at RKSC and RKSD redundancy criteria which only differ in using class-conditional distributions and total distributions.

4.1 Statistical Evaluation

We used AUC as our performance measure. Following the recommendations of [28] we did not base our statistics on performances of single folds but took averages (medians[7]) over folds.

[7] According to the central limit theorem, any sum (such as e. g. a performance benchmark), if of finite variance, of many independent identically distributed random features will converge to a Gaussian distribution. This is however not necessarily to expect for only 5 values, i. e. from 5-folds of cross-validations. After finding partly huge differences between means and medians over cross-validations, in pre-trial runs,

In table 1 redundancy and relevance combinations are compared over all classifiers, all numbers of features, and mRmRQ, mRmRD, and Hopfield. Tables 2 and 3 analyze redundancy measures and relevance measures, respectively, over all classifiers, numbers of features, mRmRQ/D, and Hopfield, and relevance or redundancy measures, respectively.

A difficulty with regard to the Greedy method is that it produces feature sets with an unpredictable number of features. We included all Greedy schemes in number-of-features specific comparison tables using a threshold of $\frac{|s_{\text{design}} - s_{\text{Greedy}}|}{s_{\text{design}}} \leq 0.1$.

We now explain the format of the result tables. The first column gives the name of the method, specified by selection scheme, redundancy, and relevance measures[8]. The second column indicates the rank of the method within methods compared in the same table. Ordering follows by mean rank of performance (third column). Median performance and interquartile range of the vector of performance scores (columns four and five) served for statistical comparisons by Friedman test and Nemenyi post-hoc test (F/N), and Wilcoxon Signed Rank Test (SR). One-to-one comparisons of methods by these statistical tests can be found in columns six and seven as win and loss scores (W/L) indicating statistical significance.

4.2 Redundancy and Relevance Measures

In table 1 you can find a ranking of RR combinations over all numbers of features and over mRmRQ/D and Hopfield.

The best combination was RVDM with FC. The table shows nearly coherent groupings by relevance measure. Everything including SU is clearly on the bottom. Also bad, but better than SU we find combinations with CC relevance. RFC and RCC redundancy seem worse than others, with RCC having greater deviation. A good redundancy measure seems to be RVDM.

In table 2 we see rankings of redundancy measures averaged (medians) over mRmRQ/D and Hopfield over all numbers of features. Here the clear winner is RJS, followed by RVDM and RST together with highly correlated RKSC and RKSD. RFC comes last, after RCC. Both had been only low correlated to the other measures (and higly negatively with each other).

A comparison of relevance measures we find in table 3. The statistics are again over mRmRQ/D and Hopfield and over all numbers of features. VDM and FC, which had been found highly correlating, are clearly the best relevance measures. CC comes before SU, which is the clear looser.

---

we decided to take the more robust median (which in case of normal distributions is equal to the arithmetic mean anyway). As for the error-bar, we plot the interquartile range (short: IQR), which is the difference between values at the first (25%) and the third quartile (75%).

[8] in the case of table 1 the average is taken over selection scheme

	index	mean rank	median	iqr	F/N W/L	SR W/L
RVDM+FC	1	7.85	0.97	0.04	17/0	25/0
RCC+VDM	2	7.99	0.97	0.03	17/0	25/1
RVDM+VDM	3	8.38	0.97	0.04	18/0	24/0
RJS+VDM	4	8.63	0.97	0.04	17/0	24/1
RJS+FC	5	9.63	0.97	0.04	14/0	19/4
RKSC+VDM	6	9.89	0.96	0.07	17/0	19/4
RST+VDM	7	10.07	0.97	0.04	16/1	20/4
RKSD+VDM	8	10.14	0.96	0.07	16/2	15/6
RFC+VDM	9	10.39	0.97	0.03	16/2	16/4
RKSC+FC	10	10.68	0.96	0.09	16/0	14/7
RKSD+FC	11	10.73	0.96	0.09	16/1	14/7
RST+FC	12	11.07	0.97	0.08	15/3	16/7
RCC+FC	13	13.72	0.96	0.07	8/10	10/12
RJS+CC	14	13.94	0.96	0.05	9/12	13/10
RFC+FC	15	13.95	0.96	0.07	9/9	9/13
RST+CC	16	14.10	0.95	0.06	8/11	11/13
RVDM+CC	17	14.52	0.96	0.06	8/12	10/15
RCC+CC	18	15.42	0.95	0.05	7/14	10/8
RKSC+CC	19	15.62	0.95	0.08	9/13	9/17
RKSD+CC	20	16.08	0.95	0.08	8/14	8/18
RFC+CC	21	17.52	0.94	0.04	7/17	7/20
RJS+SU	22	17.92	0.94	0.09	6/21	6/21
RVDM+SU	23	19.60	0.87	0.16	3/22	3/22
RST+SU	24	21.28	0.88	0.13	4/23	4/23
RKSC+SU	25	21.51	0.86	0.15	3/23	3/23
RKSD+SU	26	21.98	0.86	0.15	2/24	2/24
RFC+SU	27	26.62	0.79	0.18	0/26	1/26
RCC+SU	28	26.79	0.84	0.17	0/26	0/27

Table 1. RR Combinations over mRmRQ/D and Hopfield, and over all Numbers of Features

	index	mean rank	median	iqr	F/N W/L	SR W/L
RJS	1	2.68	0.97	0.04	5/0	6/0
RVDM	2	2.94	0.96	0.04	3/0	5/1
RST	3	3.82	0.96	0.07	2/2	3/2
RKSC	4	4.23	0.95	0.08	2/2	2/3
RKSD	5	4.38	0.95	0.08	1/3	1/4
RCC	6	4.54	0.95	0.06	0/2	0/2
RFC	7	5.40	0.95	0.05	0/4	0/5

Table 2. Redundancy over mRmRQ/D and Hopfield, and all Numbers of Features

	index	mean rank	median	iqr	F/N W/L	SR W/L
VDM	1	1.49	0.97	0.04	2/0	3/0
FC	2	1.88	0.97	0.06	2/0	2/1
CC	3	2.76	0.95	0.04	1/2	1/2
SU	4	3.86	0.86	0.12	0/3	0/3

Table 3. Relevance over mRmRQ/D and Hopfield, and all Numbers of Features

4.3 Class-Conditional Distributions

We used two redundancy criteria based on the Kolmogorov-Smirnov (KS) test, RKSC and RKSD. RKSD was computed based on the total distributions and RKSC on the class-conditional distributions, i. e.

$$RKSD(X_1, X_2) = KS(X_1, X_2) \tag{16}$$

and

$$RKSC(X_1, X_2, Y) = \frac{1}{d} \sum_{i=1}^{d} KS([X_1|Y = c_i], [X_2|Y = c_i]), \tag{17}$$

where KS refers to the p-values of the KS test.

We had introduced both RKSC and RKSD in order to test, whether it is better to use class-conditional distributions for redundancy estimation. They had a Spearman correlation coefficient of 0.96.

Table 2 shows the small difference between the two measures could have made a difference in performance with RKSC performing better than RKSD. The difference in performance is statistically significant according to the Wilcoxon test, but not significant according to the stricter Friedman and Nemenyi tests. We can conclude that estimations based on class-conditional distributions serve equal or better for redundancy measures than estimations based on the distribution totals.

5 Conclusions

In this article, we presented a framework for measuring redundancy and relevance of features and compared several measures. We present several measures of redundancy and relevance within this framework, including VDM and the fit criterion (FC) which helped us to select a feature set for our classification task. As for relevance and redundancy measures, while there cannot be any single universally best measure for all applications, we hope that our experimental comparison can give some hints as to the applicability and usefulness of some measures.

The comparison of redundancy measures and as well of relevance measures is complicated because of different scales and different levels of distinction For example, the KS-test gave very few different values, while RFC gave a broad

variety of different values. Relevance measures differ greatly with respect to the importance they assign to different features. VDM and FC, and SU and CC demonstrated large correlations ($\rho > 0.65$). Relevance measures seem to concur on the relevance on some features, however there are huge differences with respect to others. In particular, we observed that CC and SU attribute lower relevance to some Gabor filters than to some probes. RKSC and RKSD (unsurprisingly because they are so similar) were found very highly correlating with one-another ($\rho = 0.96$). Both of them also were highly correlated with RST($\rho > 0.8$). RCC correlated negatively with some measures, most markedly with RFC ($\rho = -0.61$).

As for the redundancy measures, the Jensen-Shannon Divergence, RVDM, and the sign-test were good. RFC which is based on the relevance measure FC may have been too simple. There are other options for redundancy fit criterion, for example, quantifying only the number of incorrectly classified points. Better options could also instead of binary sequences hits_{X_k} involve continuous values between 0 and 1 that express confidence of assignment.

As for symmetric uncertainty, we did not optimize the density estimation beforehand and took the most simple and straightforward means we could imagine and which worked fine for the naive Bayes. We think that this density estimation affected SU. We concede that a more careful treatment may be necessary.

Of the other relevance measures, VDM and the fit criterion were the best. CC suffered from that it favored the zero-feature. Because of the formulation, Spearman rank correlation coefficients are unsuitable for comparisons between distributions with highly unequal scales, such as the case for comparing classes (set cardinality 2) and continuous features. The Pearson correlation coefficient suffers the same weakness [29]. We expect, the Kendall rank correlation coefficient (see [30]), another much used rank correlation, to have similar problems in dealing with distributions. Other correlation measures could bring an improvement, such as possibly [31].

RVDM and RFC performed very good as unitary filters. Integration of SU makes performance degrade in many cases with a given redundancy measure when compared to other relevance measures. RCC is a bad measure for redundancy; performance was worst when using only RCC (Red:RCC) and any information helped improve performance. RKSD was also bad, RKSC slightly better. Over the different integration schemes, the measures for redundancy and relevance differed in their contribution.

We computed normalized frequencies of probes for selection based on either only relevance or only redundancy (not shown). As for relevance measures, VDM and FC came before CC and SU (which corresponds to their performance ranking). As for redundancy measures, RCC lets slip in many probes, which seems to have caused the mediocre performances with RCC redundancy. RFC and RJS also were more tolerant to probes.

References

1. Vyas, V.S., Rege, P.: Automated texture analysis with gabor filters. GVIP Journal, issue 1 **6** (2006) 35–41
2. Saeys, Y., Inza, I.n., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics **August 24** (2007)
3. Mundra, P.A., Rajapakse, J.C. In: SVM-RFE with Relevancy and Redundancy Criteria for Gene Selection. Volume 4774. Springer Berlin / Heidelberg (2007) 242–252
4. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8) (2005) 1226–1238 Member-Hanchuan Peng.
5. Duch, W., Biesiada, J.: Feature selection for high-dimensional data: A kolmogorov-smirnov correlation-based filter solution. In Kurzynski, M., Puchala, E., Wozniak, M., Zolnierek, A., eds.: Advances in Soft Computing. Springer (2005) 95–104
6. Novovicová, J., Malík, A., Pudil, P.: Feature selection using improved mutual information for text classification. In: International Workshop on Structural and Syntactic Pattern Recognition. (2004)
7. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res. **5** (2004) 1205–1224
8. Knijnenburg, T.A.: Selecting relevant and non-relevant features in microarray classification applications. Master's thesis, Delft Technical University, Faculty of Electrical Engineering, 2628 CD Delft (2004)
9. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. In: Second IEEE Computational Systems Bioinformatics Conference. (2003) 523–529
10. Zhang, Y., Callan, J., Minka, T.: Novelty and redundancy detection in adaptive filtering. In: SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (2002) 81–88
11. Zhou, J., Peng, H.: Automatic recognition and annotation of gene expression patterns of fly embryos. Bioinformatics **23**(5) (2007) 589–596
12. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. (2005)
13. Liu, X., Krishnan, A., Mondry, A.: An entropy-based gene selection method for cancer classification using microarray data. BMC Bioinformatics **6** (2005)
14. Yu, L., Liu, H.: Feature selection for high-dimensional data: A fast correlation-based filter solution. In: ICML. (2003) 856–863
15. Conover, W., Iman, R.: Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics. AM. STAT. **35**(3) (1981) 124–129
16. Wu, G., Twomey, S., Thiers, R.: Statistical Evaluation of Method-Comparison Data. Clinical Chemistry **21**(3) (1975) 315–320
17. Stanfill, C., Waltz, D.: Toward memory-based reasoning. Communications of the ACM **29**(12) (1986) 1213–1228
18. Wilson, D.R., Martinez, T.R.: Improved heterogeneous distance functions. Journal of Artificial Intelligence Research 6 **6** (1997) 1–34
19. Payne, T.R., Edwards, P.: Implicit feature selection with the value difference metric. In: European Conference on Artificial Intelligence, - (1998) 450–454
20. Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory **37**(1) (Jan. 1991) 145–51

21. Auffarth, B., López-Sánchez, M., Cerquides, J. In: Hopfield Networks in Relevance and Redundancy Feature Selection Applied to Classification of Biomedical High-Resolution Micro-CT Images. Petra Perner (July 2008)
22. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001) Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
23. Burt, P.J., Adelson, E.H.: The laplacian pyramid as a compact image code. IEEE Trans. Communications **31** (1983) 532–540
24. Kovesi, P.D.: Edges are not just steps. Proceedings of the Fifth Asian Conference on Computer Vision (January 2002) 822–827
25. Reinagel, P., Zador, A.: Natural scene statistics at center of gaze. Network: Comp. Neural Syst. **10** (1999) 341–350
26. Einhäuser, W., Kruse, W., Hoffman, K.P., König, P.: Differences of monkey and human overt attention under natural conditions. Vision Research **46(8-9)** (2006) 1194–1209
27. Auffarth, B.: Classification of biomedical high-resolution micro-ct images for direct volume rendering. Master's thesis, University of Barcelona, Barcelona, Spain (2007)
28. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research **7** (2006) 1–30
29. Bollen, K., Bollen, K.: Structural equations with latent variables. Wiley New York (1989)
30. Abdi, H.: The Kendall Rank Correlation Coefficient, edited by NJ Salkind. Encyclopedia of Measurement and Statistics (2007)
31. Yilmaz, E., Aslam, J., Robertson, S.: A new rank correlation coefficient for information retrieval. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, ACM New York, NY, USA (2008) 587–594