2009 Ninth International Conference on Intelligent Systems Design and Applications

Empirical Study of Individual Feature Evaluators and Cutting Criteria for Feature Selection in Classification

Antonio Arauzo-Azofra*, José L. Aznarte M.† and José M. Benítez‡
*Area of Project Engineering, University of Cordoba, Cordoba, Spain
Email: arauzo(at)uco.es
†Centre for Energy and Processes, Ecole des Mines de Paris, Sophia Antipolis, France
Email: jose-luis.aznarte(at)mines-paristech.fr
‡Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain
Email: J.M.Benitez(at)decsai.ugr.es

***Abstract*—The use of feature selection can improve accuracy, efficiency, applicability and understandability of a learning process and its resulting model. For this reason, many methods of automatic feature selection have been developed. By using a modularization of feature selection process, this paper evaluates a wide spectrum of these methods. The methods considered are created by combination of different selection criteria and individual feature evaluation modules. These methods are commonly used because of their low running time. After carrying out a thorough empirical study the most interesting methods are identified and some recommendations about which feature selection method should be used under different conditions are provided.**

***Keywords*-feature selection; feature evaluation; attribute evaluation; classification**

I. Introduction

The task of a classifier is to use feature vectors to assign the represented object to a category or class [1]. Feature selection helps us to focus the attention of a classification algorithm on those features that are the most relevant to predict the class. Theoretically, if the full statistical distribution were known, using more features could only improve results. However, in practical learning scenarios, it may be better to use a reduced set of features [2].

Sometimes, a large number of features in the input of induction algorithms may turn them inefficient as memory and/or time consumers, even turning them inapplicable. Besides, irrelevant data may confuse algorithms leading them to reach false conclusions, and hence producing worse results. Other advantages of using feature selection may be improving understandability and lowering costs of data acquisition and handling. Because of all these advantages, feature selection has attracted much attention within the Machine Learning and Data Mining communities and many methods have been developed [3], [4], [5], [6] with diverse applications.

According to the different parts identified in feature selection methods [3], [7], [8], its process can be modularized [9] as shown in figure 1. With this modularization almost every feature selection method can be characterized through the evaluation function and search strategy employed.

One widely used group of feature selection methods is formed by those using the evaluation of individual features (the lower evaluation module in figure 1) together with a cutting criterion to select the features (this can be seen as a simple search module). The goal of this paper is to carry out an extensive and rigorous empirical evaluation of these feature selection methods applied in classification.

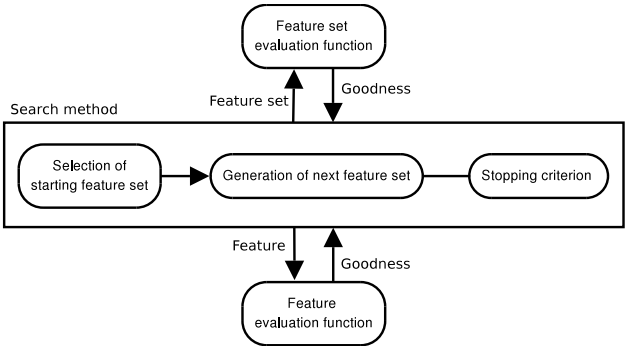

Figure 1. Feature selection modularized

II. Feature Selection Methods

The methods considered utilize evaluation functions that assign an evaluation value to each feature. After the evaluation process, the features with a higher evaluation are chosen, but some method is necessary to determine how many features are selected. There are many possible cutting criteria to perform this task. However, up to our knowledge, there is no study in order to decide which cutting criterion is the most appropriate. Many proposals simply establish that some threshold should be chosen and leave the choice to the practitioner.

This work considers a set of feature selection methods that are created by combination of:

1) a feature evaluation measure to assign preference values to features and

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.175

2) a cutting criterion to choose the number of features selected

The description of the five measures considered follows:

Mutual information (info),
: also known as information gain, measures the quantity of information that a feature gives about the class. It comes from the Information Theory of Shannon [10] and it is defined as the difference between the entropy of the class and the entropy of the class conditioned to know the evaluated feature.

Gain ratio (gain)
: is defined as the ratio between information gain and the entropy of the feature. In this way, this measure avoids favoring features with more values, which is the natural behavior of previous measure. This measure was used by Quinlan in his C4.5 algorithm [11].

Gini index (gini)
: can be seen as the probability of two instances randomly chosen having a different class. It was used by Breiman [12] to generate classification trees.

Relief-F (reli)
: is an extension of the original Relief [13] developed by Kononenko [14]. It can handle discrete and continuous attributes, as well as null values. Despite evaluating individual features, Relief takes into account relations among features. This makes Relief-F to perform very well, becoming well known and very commonly used in feature selection.

Relevance (rele)
: is a measure that discriminates between attributes on the basis of their potential value in the formation of decision rules [15].

In this study, we intend to apply general cutting criteria. They have been designed to be used with any measure in any data set. The description of the six cutting methods chosen follows.

N best (n)
: simply selects a fixed number of features.

Fraction (p)
: selects a fraction, given as a percentage, of the total number of available features.

Threshold (t)
: selects the features whose evaluation is over a user given threshold.

Threshold given as a fraction (pm)
: selects the features whose evaluation is over a threshold, where this threshold is given as a fraction of the range of evaluation function.

Difference (d)
: selects features, starting from the one with greater evaluation and following the sorted list of features, until evaluation difference is over a threshold.

Slope (s),
: on the sorted list of features, selects best features until the slope to the next feature is over a threshold.

All combinations of the feature evaluation measures and cutting criteria considered are feasible, so 30 methods will be evaluated.

III. Empirical methodology

With the goal stated in the introduction in mind, we designed and conducted an extensive and rigorous empirical study. In this section, we provided a detailed description of the experimental method followed.

A. Experimental design

The main measures to be taken into account when evaluating a feature selection method are accuracy and feature reduction.

In our classification task, there are three main factors:

1) Feature selection method, with measure and cutting criterion as subfactors
2) Learning algorithm that generates the classifier
3) Classification problem represented in a data set

The goal of this work is to compare feature selection methods taking into account all factors, so a complete experimental setup has been used. In this setup, the number of independent experiments is the number of the possible combinations of the three factors above.

In order to get reliable estimates for classification accuracy on each classification task, every experiment has been performed using 10 fold cross-validation. Any result shown is always the average of the 10 folds.

The significance of results is assessed using statistical tests. To choose the right test, two features of results must be noted. First, classification rates among data sets are not commensurable, as results on different data sets are not comparable for a given classifier. And second, many methods are compared at the same time. Since we are performing multiple comparisons, we can not simply repeat –so many times– the tests designed for a pair of variables, as the number null hypotheses rejected by random chance will become high. Following the methodology recommended by Demsar [16] for this type of comparisons, we have used Iman Davenport and Nemenyi statistical tests. A detailed description of these tests can be found in Zar's book [17].

B. Data sets

In order to include a wide range of classification problems, publicly available repositories [18] [19] [20] [21] have been explored, seeking for representative problems with different properties (discrete and continuous data, different number of classes, features, examples, and unknown values). Finally,

the following 36 diverse data sets have been used: adult, anneal, audiology, balance-scale, breast-cancer, bupa, car, credit, echocardiogram, horse-colic, house-votes84, ionosphere, iris, labor-neg, led24, lenses, lung-cancer, lymphography, mushrooms, parity3+3, pima, post-operative, primary-tumor, promoters, saheart, shuttle-landing-control, soybean, splice, tic-tac-toe, vehicle, vowel, wdbc, wine, yeast, yeast-class-RPR, and zoo.

C. Classifiers

In order to estimate the quality of feature selection performed by each method, the selected features are tested in a complete learning scenario of classification problems. The following well known learning methods [1] are considered. To set up parameters of learning methods, preliminary experiments with different parameter values were performed on the data sets.

- Naive-Bayes (with LOESS for continuous data) [15] (Nbayes), a simple method that establishes a base on the minimal performance that other more elaborated methods should improve on.
- k Nearest Neighbors [15] (kNN). This method has been considered as a representant of those methods that use distances in classification. After the preliminary experiments, the value $k = 15$ was chosen as a value large enough to get good results in all considered data sets.
- Classification trees (C45). We intend this classifier to represent tree and rule based classifiers in our experiments. C4.5 [11] is well known and commonly used to evaluate feature selectors.
- Artificial Neural Networks [22] (ANN). As a representation of ANN in classification we have chosen the well known multilayered perceptron with one hidden layer. The number of nodes in the hidden layer is adjusted to the average between the number of inputs and outputs. The network will have one output per class, and the class is decided by the output with the highest value. The training algorithm is standard back-propagation with learning rate of 0.1 and 500 learning cycles.

D. Data transformations

Some feature selection methods require certain conditions on data. Consequently, data are transformed just for these feature selection methods. After feature selection, original data are passed to the learning methods.

When necessary features were discretized using equal frequency intervals. When continuous features were required discrete features were translated to equidistant points in $[0, 1]$. For those methods that could not cope with null or unknown values, these values were replaced by the average or the most frequent value on discrete features.

E. Development and running environment

The software used for learning methods has been Orange component-based data mining software [15], except for artificial neural networks where OrangeSNNS [23] was used. The feature selection methods have been coded using the Python programming language.

F. Parameters of feature selection methods

In order to compare the methods, we want completely determined methods with fixed parameters. All evaluation functions are parameter free except Relief-F. Based on Relief-F analysis [24] and some preliminary experiments, the number of neighbors to search is set to 6, and the number of instances to sample is set to 100.

For each cutting criterion, some reasonable values of the parameters have been tested. The finally chosen value is the one which has lead to best average ranking in accuracy over all data sets and measures. Cutting criteria are refered in experiment results by the abreviation formed with its name as given in section 2 and the value of its parameter (n17, p0.8, t0.1, pm0.8, d0.2, and s1.5).

IV. Experimental Results

The experimental results are extense so, just the most relevant will be commented organized in three parts. First, a comparison of the evaluation functions. Second, a comparison of cutting criteria and, finally, the comparison of the composed methods.

A. Comparing evaluation functions

For every cutting criterion, all feature evaluation measures have been compared. Figure 2 shows the comparison of feature evaluation functions using the *t0.1* cutting criterion. On these figures, the abscissa axis represents the ranking of each measure in relation with the others. The ranking belongs to the interval $[1, n]$ when comparing n measures (the lower the ranking the better accuracy). The value shown for each measure is the average ranking over the 36 data sets.

The rows of figure 2 show the results for each of the four considered learning methods. In this way, we can compare the effect of feature selection on each learner and we assure independence for the application of statistical tests. The rectangle shows Nemenyi critical distance from the best method, at significance level of $p = 0.05$. Those methods outside the rectangle can be considered to obtain a worse accuracy. Besides, if the rectangle is slipped, all methods that can be separated by the rectangle (lying one on each side) have a significative difference on accuracy.

Lower figure 2 shows the same comparison but on the number of features selected (the lower the ranking the greater the reduction). Only one reduction is shown as the feature selection is the same for all the classifiers.

From these figures about t0.1 comparison, we can see Relief (*reli*) leading on accuracy, though significative difference

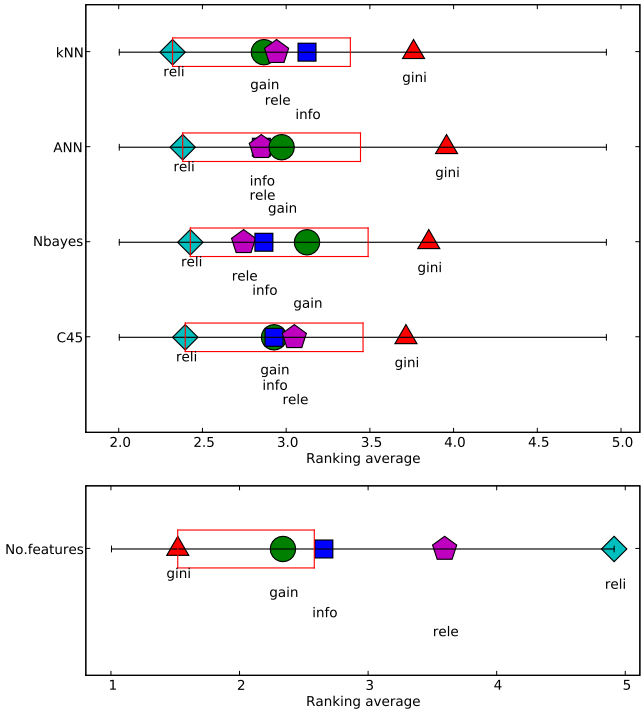

Figure 2. Comparison of feature evaluation functions with cutting criterion t0.1

can only be found with Gini ratio (*gini*). However, Relief is the evaluation function that achieves worst feature reduction with this cutting criterion, while *gini* is the measure that offers the greatest reductions. Having in mind both objectives, probably, a good choice is using information gain ratio (*gain*) because, while applying the second greatest reduction with no significative difference with the first (*gini*), *gain* achieves good accuracy (near the first with no significative difference).

Comparing measures with other cutting criteria, the following facts can be observed. For space reasons, all these figures can not be included. *n17* and *p0.8* apply equal reduction for all measures and differences on accuracy are not significant.

As *t0.1*, all the remaining cutting criteria perform lower or greater feature reduction depending on the measure value. No significative differences have been found with them. Anyway, the differences coming from experiments are commented now. Using *pm0.8* cutting criterion, *gain*, *info* and *gini* stand out on accuracy and reduction simultaneously. However, *reli* seems not appropriate to be used with this cutting criterion as it gets worst accuracy and reduction results. With *d0.2*, rules relevance (*rele*) performs best with similar accuracy to the others and with best reduction. Relief (*reli*) works better than with pm0.8 as improves on reduction, but it is still the worst on accuracy. Gini index obtains similar accuracy to the other methods, but being the worst on reduction. Using *s1.5* the differences are smaller.

B. Comparing cutting criteria

On figure 3, cutting criterion methods are compared when using *gain* measure. While applying the greatest reductions, *pm0.8* and *s1.5* criteria get the worst results on accuracy with a significative difference. The *t0.1* cutting criterion obtains intermediate results on both accuracy and reduction, where significative differences have not been detected with the first method on both concepts. The rest of methods depend on the learner being applied after them. Considering feature reduction, *p0.8* get better reduction than *n17* and *d0.2*. On kNN and ANN, *n17* leads accuracy results. On C4.5, *n17* and *p0.8* lead accuracy results, while on NBayes, all of them obtain pretty similar results.

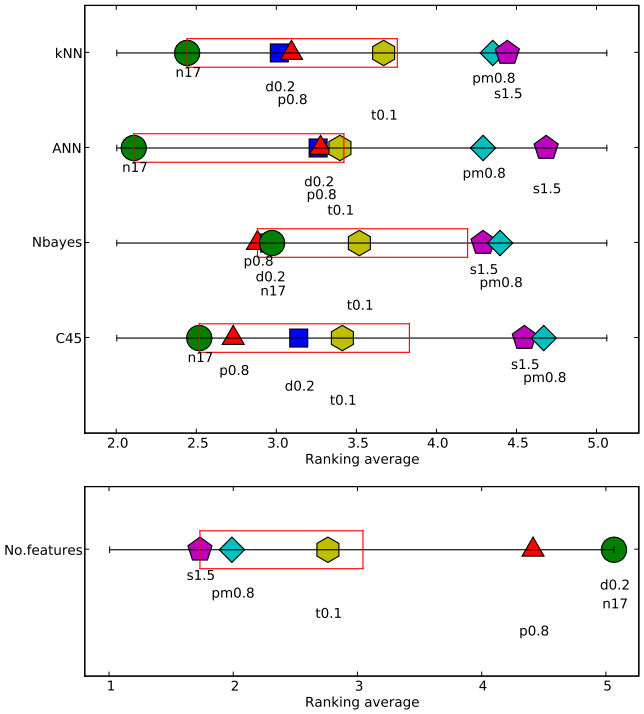

Figure 3. Comparison of cutting criteria with gain evaluation function

Comparing cutting criteria with other measures, the following facts can be observed. *Reli* results are similar to those of *gain*, but with *t0.1* becoming the worst on reduction and improving a bit on accuracy. Considering the *gini* measure, two groups of cutting criterion methods can be clearly distinguished, as a significative distance separates them. The first group is composed by those that achieve higher accuracy with lower reductions. The three cutting criteria (*n17*, *p0.8*, and *d0.2*) obtain very similar accuracy, achieving greater reductions in this order: *p0.8*, *n17* and *d0.2*. The other group is formed by those methods that apply higher reductions and achieve lower accuracy (*pm0.8*, *t0.1*, and *s1.5*). Using the *rele* measure, the two cutting criteria

that have achieved worst accuracy with greater reduction are more separated of each other —near to a significative distance at 0.05, but lower— than in all the other measures. In this way, *p0.8* get better results on accuracy while keeping the same distance to *s1.5* on reduction than in all the other measures. On NBayes, all ahead criteria (*n17*, *p0.8*, and *d0.2*) get practically the same results on accuracy, while on C4.5, *p0.8* leads accuracy with greater reduction and *n17* leads on kNN and ANN.

It can be concluded that the cutting criterion to use depends on the measure and the learner. Despite our efforts to normalize measures and set cutting criterion parameters that performs best with all measures, the results vary among measures and no criterion can be generally recommended.

C. Global comparison of feature selection methods

The measures and cutting criteria have been compared above varying just one of them independently. All the 30 combinations of measures and cutting criteria considered can not be displayed clearly on a figure like the previous ones. Neither a table with all the results (non aggregated) can be interpreted. For this reason, a table with ranking from the global comparison and some different figures are shown here.

The results of comparing all methods at the same time are shown in table I. The values are the average ranking over the 36 data sets. This table also considers not applying feature selection at all, this is identified as *NoFS* method. Data for each learner is ranked independently of the others

Looking at this table, we can see that *n17-info*, *n17-gain*, *n17-gini* improve accuracy results from not applying feature selection (except for NBayes). This means that applying feature selection, even with these simple methods can improve results while reducing data used by the learner. The method *p0.8-gini* has provided greater reduction without without a great accuracy loss.

The best method for each learner is marked with a boldface type in table I. These are *n17-gain* for ANN, *p0.8-reli* for C4.5, *n17-info* for kNN, and *d0.2-info* for NBayes. While they do not reduce much the number of features, all of them reduce features and perform better than *NoFS*. On the other side, the method that applies the greatest reduction is *t0.1-gini*, but its results on accuracy —though not the worst— do not give much confidence about its application to other problems.

Feature selection method	*Accuracy (avg. rank)*				Feat. red.
	C45	*Nbay.*	*ANN*	*knn*	
NoFS-	12.88	12.61	10.02	12.35	26.91
d0.2-info	14.12	**11.58**	14.26	11.45	23.58
d0.2-gain	14.61	13.14	14.64	13.80	23.09
d0.2-gini	12.80	12.61	12.98	12.35	26.91
d0.2-reli	17.35	16.41	16.11	16.03	21.21
d0.2-rele	13.67	12.39	14.55	14.15	21.71
n17-info	10.89	13.24	9.65	**10.39**	22.98
n17-gain	11.27	13.39	**9.18**	10.62	22.98
n17-gini	11.03	13.17	9.98	11.17	22.98
n17-reli	12.33	14.62	10.59	12.70	22.98
n17-rele	11.65	13.27	10.32	11.27	22.98
p0.8-info	10.09	15.38	12.65	12.42	18.73
p0.8-gain	11.79	14.09	14.73	13.26	18.73
p0.8-gini	10.83	13.53	12.50	12.35	18.73
p0.8-reli	**9.35**	16.08	14.56	14.45	18.73
p0.8-rele	10.73	12.94	13.65	13.35	18.73
pm0.8-info	21.79	19.35	19.55	20.73	7.17
pm0.8-gain	21.85	19.88	20.42	20.33	7.35
pm0.8-gini	21.42	18.50	18.70	19.38	7.59
pm0.8-reli	20.02	20.53	22.00	20.32	9.18
pm0.8-rele	19.76	18.35	18.95	19.14	8.65
s1.5-info	22.74	19.61	22.03	21.50	6.00
s1.5-gain	22.17	19.97	21.86	20.94	6.48
s1.5-gini	22.71	20.08	22.55	21.62	5.91
s1.5-reli	21.26	21.21	23.02	22.41	8.06
s1.5-rele	23.18	21.70	22.80	22.20	7.08
t0.1-info	17.00	15.21	16.17	17.18	11.33
t0.1-gain	15.15	16.09	15.47	16.65	10.36
t0.1-gini	21.39	20.62	22.97	21.83	**5.59**
t0.1-reli	12.65	12.61	13.23	12.35	26.91
t0.1-rele	17.52	13.86	15.92	17.30	16.36

Table I
GLOBAL COMPARISON OF ALL FEATURE SELECTION METHODS CONSIDERED

V. CONCLUSIONS

In the field of classification problems, a rigorous empirical study on individual feature evaluation measures and cutting criterion methods has been presented. All methods created by combination of the evaluation functions and the cutting criteria chosen are explored with a state of the art experimental design.

While reducing the number of features used, some of the feature selection methods evaluated improve results in most of the problems considered. This confirms the usefulness of feature selection.

A contraposition of accuracy and feature reduction is detected, showing that methods performing greater reductions start losing relevant features leading to worse accuracy results. For this reason, a single method can not be recommended for all situations, the table with methods ranked on accuracy and reduction for each learner can be used as a guide.

About evaluation functions, in general, those based on information theory reached better accuracy results, while for C4.5 learner Relief was the best measure.

No cutting criterion can be generally recommended. Those independent from measure reach better accuracy results, while the others reach higher feature reductions.

Results vary among learners, having different feature selection methods that perform well with each of them. Altough, this would recommend using the wrapper approach, this approach could not compete with individual feature evaluation methods in computing time nor in applicability

with a large number of features. On that kind of problems, some of the evaluated methods are more appropiate.

REFERENCES

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2001.

[2] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[3] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, no. 1-4, pp. 131–156, 1997.

[4] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Springer, 1998.

[5] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, March 2005.

[6] K. Thangavela and A. Pethalakshmi, "Dimensionality reduction based on rough set theory: A review," *Applied Soft Computing*, vol. 9, no. 1, pp. 1–12, Jan 2008.

[7] A. Jain and D. Zongker, "Feature selection: Evaluation, application and small sample performance," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, February 1997.

[8] P. Langley, "Selection of relevant features in machine learning," in *Procedings of the AAAI Fall Symposium on Relevance*. New Orleans, LA, USA: AAAI Press, 1994, pp. 1–5.

[9] A. Arauzo-Azofra, J. M. Benitez, and J. L. Castro, "Consistency measures for feature selection," *Journal of Intelligent Information Systems*, vol. 30, no. 3, pp. 273–292, June 2008.

[10] T. M. Cover and J. A. Thomas, *Elements of Information theory*. Wiley-Interscience, 1991.

[11] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[12] L. Breiman, Ed., *Classification and regression trees*. Chapman & Hall, 1998.

[13] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proceedings of the Ninth International Conference on Machine Learning*. Aberdeen, Scotland: Morgan Kaufmann, 1992, pp. 249–256.

[14] I. Kononenko, "Estimating attributes: Analysis and extensions of RELIEF," in *European Conference on Machine Learning*, 1994, pp. 171–182. [Online]. Available: citeseer.nj.nec.com/kononenko94estimating.html

[15] J. Demsar and B. Zupan, "Orange: From experimental machine learning to interactive data mining," (White paper) http://www.ailab.si/orange, 2004.

[16] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[17] J. H. Zar, *Biostatistical Analysis*, 4th ed. New Jersey (US): Prentice-Hall, 1999.

[18] S. Hettich and S. D. Bay, "Uci machine learning repository," http://archive.ics.uci.edu/ml/, 2008.

[19] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag, 2001.

[20] S. Graphics, "Datasets," http://www.sgi.com/tech/mlc/db/, 2006.

[21] U. of Toronto, "Data for evaluating learning in valid experiments," http://www.cs.utoronto.ca/~delve/, 2003.

[22] S. Haykin, *Neural Networks. A comprehensive foundation*. Prentice-Hall, 1999.

[23] A. Arauzo-Azofra, "Orange snns module," http://ax5.com/antonio/orangesnns/, 2006.

[24] M. Robnik-Sikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," *Machine Learning*, vol. 53, pp. 23–69, 2003.