Comparison of feature ranking methods based on information entropy.

Włodzisław Duch
Department of Informatics, Nicholaus Copernicus University,
Grudziądzka 5, Toruń, Poland,
and School of Computer Engineering,
Nanyang Technological University, Singapore.
www.phys.uni.torun.pl/∼duch

Tadeusz Wieczorek, Jacek Biesiada, Marcin Blachnik
Division of Computer Methods,
Department of Electrotechnology,
The Silesian University of Technology,
ul. Krasińskiego 8, 40-019 Katowice, Poland.
marcinblachnik@poczta.onet.pl

***Abstract*— A comparison between five feature ranking methods based on entropy is presented on artificial and real datasets. Feature ranking method using χ^2 statistics gives results that are very similar to the entropy-based methods. The quality of feature rankings obtained by these methods is evaluated using the decision tree and the nearest neighbor classifier with growing number of most important features. Significant differences are found in some cases, but there is no single best index that works best for all data and all classifiers. Therefore to be sure that a subset of features giving highest accuracy has been selected requires the use of many different indices.**

I. Introduction

Feature selection is a problem that has to be addressed in many areas, especially in bioinformatics, text analysis, object recognition or in modeling of complex technological processes. Bioinformatics datasets frequently contain thousands, or even hundreds of thousands of features. Good examples of several highly-dimensional datasets have been provided in the NIPS 2003 challenge on feature extraction [1]. All features may be important for some problems, but from a specific point of view, frequently related to recognition of some target concepts, only a small subset of features is usually relevant. Many solutions in highly dimensional feature spaces with limited amount of available data exist due to accidental correlations between the target concept and various ways of partitioning the data, making these solutions worthless. To deal with such problems dimensionality of the feature space has to be reduced first. This may be done by selecting a subset of relevant features from the total number of features, or by ranking these features and selecting the most important ones.

Many feature selection and feature ranking methods have been proposed in the literature (see for example [2], [9], [11], [14]). Ranking of features determines the importance of any individual feature, neglecting their possible interactions. Ranking methods are based on statistics, information theory, or on some functions of classifier's outputs [5]. In this paper a few entropy based methods are compared. Algorithms for feature selection fall into two broad categories: wrappers that use the learning algorithm itself to evaluate the usefulness of features, and filters that evaluate features according to heuristics based on general characteristics of the data [4], [7]. Some classification algorithms have inherent ability to focus on relevant features and ignore irrelevant ones. Decision trees are primary example of a class of such algorithms [1], [15], but also multi-layer perceptron (MLP) neural networks with strong regularization of the input layer may turn off the irrelevant features in an automatic way [3]. Such methods may also benefit from independent feature selection. On the other hand some algorithms have no provisions for feature selection. The k-nearest neighbors algorithms (k-NN) are one family of such methods that classify novel examples by retrieving the nearest stored training example, relying on independent feature selection methods.

Wrapper methods employ statistical re-sampling technique (such as crossvalidation) using the actual target learning algorithm to estimate the accuracy of feature subsets. This approach has proved useful, but is very slow to execute because the learning algorithm is called repeatedly. For this reason wrappers do not scale well to large datasets containing many features. Filter methods, on the other hand, operate independently of any learning algorithm and undesirable features are filtered out of the data. Filters typically make use of all the available training data when selecting a subset of features. Other filter methods attempt to rank features according to a relevancy score.

The next section discusses entropy based feature ranking indices. Section 3 describes the datasets used in experiments. Section 4 presents experimental results comparing five entropy based methods to a method introduced here, based on a simple statistical measure χ^2 of independence of feature/class distributions. A summary of the results and plans for future work is given in the last section.

II. Theoretical framework

A typical ranking process consists of four steps:

1) Initialize set $\mathcal{F}$ to the whole set of p features. $\mathcal{S}$ is an empty set.
2) For all features $f \in \mathcal{F}$ compute $J(f)$-coefficient.
3) Find feature f that maximizes $J(f)$ and move it to $\mathcal{S} \leftarrow \mathcal{S} \cup \{f\}, \mathcal{F} \in \mathcal{F} \backslash \{f\}$
4) Repeat until the cardinal of $\mathcal{S}$ is p.

[1] http://clopinet.com/isabelle/NIPS2003/

where $J(f)$ is a criterion function (different for any ranking algorithm) which gives a measure of dependency between features (f) and classes (C).

First ranking algorithm uses normalized information gain, called the asymmetric dependency coefficient (ADC) [14]:

$$ADC(C,f) = \frac{MI\,(C,f)}{H\,(C)} \tag{1}$$

where for K classes information entropy $H(C)$ and $H(f)$, and mutual information $MI(C,f)$ between C and f is defined according to Shanonn [13] as:

$$\begin{aligned} H(C) = & \quad -\sum_{i=1}^{K} p(C_i)\lg_2 p(C_i) \\ H(f) = & \quad -\sum_{x} p(f=x)\lg_2 p(f=x) \\ MI(C,f) = & \quad -H(C,f)+H(C)+H(f) \end{aligned} \tag{2}$$

Here the sum over x is used only for features f that take discrete values, for continuous features it should be replaced by an integral or discretization should be performed first to estimate probabilities $p(f=x)$.

The ranking algorithm proposed by Setiono [11] uses a normalized gain ratio:

$$U_S(C,f) \; = \; \frac{MI(C,f)}{H(f)} \tag{3}$$

Another normalization may be used to calculate infromation gain for class-feature entropy :

$$U_H(C,f) = \frac{MI(C,f)}{H(f,C)} \tag{4}$$

where $H(f,C)$ is the joint entropy of f and C variables.

Mantaras [9] has proposed an interesting criterion D_{ML} which fulfills all axioms of distance, that may be defined by:

$$D_{ML}(f_i,C) \; = \; H(f_i|C) \; + \; H(C|f_i) \tag{5}$$

where $H(f_i|C)$ and $H(C|f_i)$ is the conditional entropy defined by Shanonn [13] as $H(X|Y) = H(X,Y) - H(Y)$.

Weighted joint entropy index introduced by Chi [2] is defined as:

$$Ch(f) = -\sum_{k=1}^{N} p(f=x_k)\sum_{i=1}^{K} p(f=x_k,C_i)\lg_2 p(f=x_k,C_i) \tag{6}$$

An alternative statistical measure of the dependence between two random variables – in this case the relationship between the value of a feature and the class – may be based on the χ^2 statistics. The χ^2 coefficient is given by:

$$\chi^2(f,C) = \sum_{ij}\frac{(p(f=X_j,C_i)-p(f=x_j)*p(C_i))^2}{p(f=x_j)*p_(C_i)} \tag{7}$$

where $p(\cdot)$ are probabilities. Large values of χ^2 signify strong correlation between feature values and class labels, and therefore may be used for ranking features. The χ^2 statistics has been previously used in the discretization process by Setiono and Liu [12].

Although many other coefficients measuring similarity of distributions may be introduced (for example, correlation-based coefficients [6]) it is not clear if there is any difference between them in practice. To answer this questions computational experiments described below were performed.

III. Datasets used for testing

Artificial datasets called “Gauss1" and “Gauss2" have been generated by Duch *et al.* [5] to compare different feature ranking and feature selection methods on data for which the importance of feature is known. Two real datasets for tests were used, the “hypothyroid" and “abalone" data, both taken from the UCI repository of machine learning databases [10].

A. Artificial data

* Gauss1,Gauss2
These datasets have four and eight features, respectively. In the first dataset four Gaussian functions with unit dispersion have been used to generate vectors in 4 dimensions, each Gaussian cluster representing a separate class. The first Gaussian is centered at (0,0,0,0), the next at $a(1,1/2,1/3,1/4), 2a(1,1/2,1/3,1/4)$, $3a(1,1/2,1/3,1/4)$, respectively (a is a constant). The dataset contains 4000 vectors, 1000 per each class. In this case the ideal ranking should give the following order: $x_1 > x_2 > x_3 > x_4$.
The second dataset (Gauss2) is an extension of the first, containing eight features. Additional 4 linearly dependent features have been created by taking $x_{i+4} \; = \; 2x_i + \epsilon$, where ϵ is a uniform noise with unit variance. In this case the ideal ranking should give the following order: $x_1 > x_5 > x_2 > x_6 > x_3 > x_7 > x_4 > x_8$.

B. Real data

* Hypothyroid
This data contains results from real medical screening tests for the thyroid problems. The class distribution is about 92.5% normal, 5% of the primary hypothyroid and 2.5% of the compensated hypothyroid type. The data offers a good mixture of nominal (15) and continuous (6) features. A total of 3772 cases are given for training (results from one year) and 3428 cases for testing (results from the next year of data collection).

* Abalone
The age of abalone molluscs should be predicted from

their physical measurements. This is essentially an approximation problem, but because the age is quantized into 29 bins (the number of rings, equivalent to the age in full years) it may be treated as a classification problem with a large number of classes. 4177 cases with 8 features are given, including one nominal feature, six continuous measurement values, and feature number one with values created randomly in the $[0, 1]$ range.

IV. Experiments and results

Four datasets described above have been used in numerical experiments. In each case six methods of feature ranking (Eq.1 to Eq.7) have been applied. Because the datasets have both discrete and continuous features different discretization procedures have been used first. Continuous features have been discretized in the simplest possible way: ranges of each of the features have been divided into 16 to 32 intervals of equal width. Although much better results may be expected with more sophisticated discretization [5] this is sufficient for relative comparison of entropy based methods. This has been verified using the artificial datasets Gauss1 and Gauss2. All six ranking algorithms gave optimal results in agreement with theoretical predictions, so in tests on artificial data of this type they all seem to be equivalent.

A. *Results for hypothyroid dataset*

Hypothyroid dataset has a large training and test part. 21 features are given, 15 binary and 6 continuous features (no. 1, 17, 18, 19, 20, 21) obtained from medical tests. Discretization of these 6 features based on 16 or 32 intervals does not change results. The ranking of features obtained for the training data by all six $J(f)$ indices is presented in Tab. 1. The first 4 features (all continuous) are consistently ranked as the top, although features 21 and 17 are reversed in some rankings. Significant differences are observed in the order of the remaining features.

For each ranking method investigation of classification accuracy on the test data as a function of the n best features has been done. Two classifiers were used: the nearest neighbor (as implemented in the GhostMiner package [2]) and the C4.5 decision tree [15], as implemented in Weka [3]. Both of these classifiers give deterministic results, simplifying the comparison (in contrast to these methods, neural classifiers give slightly different results after each restart, therefore averaging and variance of results should be taken into account).

Classification results are presented in Fig.1. Feature number 17, predicted by the U_S, U_H and D_{ML} indices, is definitely more important than 21. Up to 4 features these 3 methods give similar results, but then only U_S index provides important feature, all other indices leading to significant degradation of results. The last 6 features evidently confuse the nearest neighbor classifier, therefore they should always be deleted in this type of methods.

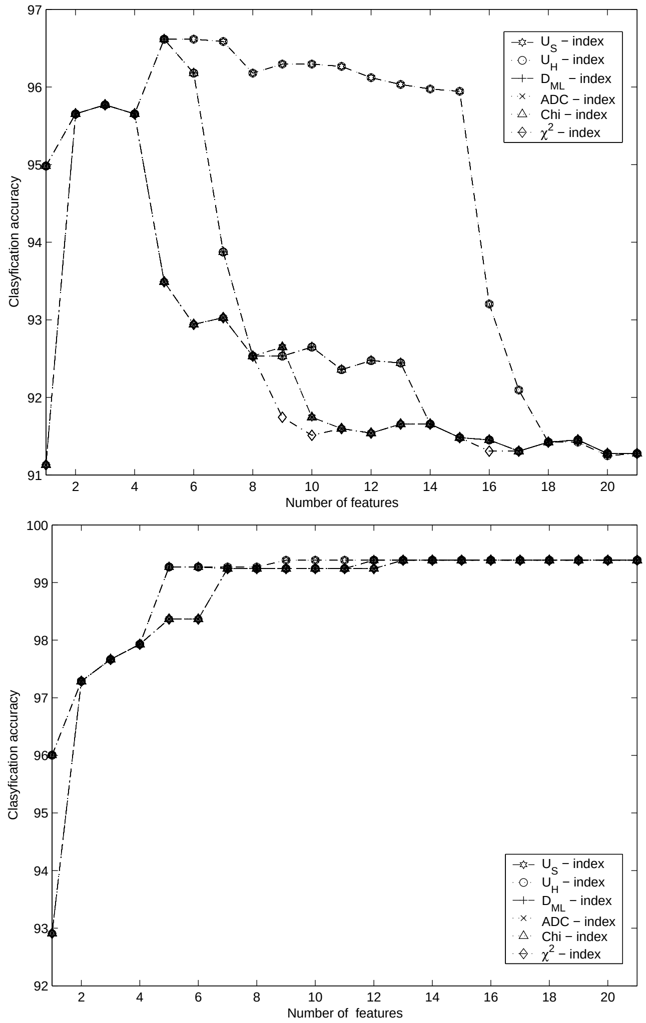

Fig. 1. Classification accuracy for the hypothyroid dataset: upper figure – 1NN classifier, lower – C4.5.

Also in case of the C4.5 decision tree classifier selection of feature 21 as the first feature gives poor results, and the peak performance is reached for 5 features. Because C 4.5 removes less useful features automatically pruning its tree accuracy does not drop but stays at the peak level as long as all important features are included. Unfortunately this time U_S index selects suboptimal feature number 3 and 7, while other indices opt for 1 and 20 and reach higher accuracy. Still U_S index is the first to add both features 10 and 8, reaching highest performance with 9 features. Thus although there is no clear overall winner, different classification methods may show some preferences, normalized information gain index U_S performs very well.

B. *Results for the abalone datasets*

Similar calculations were performed for the abalone dataset. First the ranking algorithms were applied to the whole dataset, and since there is no test datasets classification accuracy was estimated using ten-fold crossvalidation. For our purpose – comparing different ranking methods – this approach is

[2] http://www.fqspl.com.pl

[3] http://www.cw.waikato.ac.nz/ ml/weka/

TABLE I

RESULTS OF FEATURE RANKING ON THE HYPOTHYROID DATA; SEE DESCRIPTION IN THE TEXT.

Method	**Most – Least Important**																				
ADC index, Eq. 1	21	17	19	18	1	20	3	10	16	2	6	7	8	13	5	4	11	12	14	9	15
U_S index, Eq. 3	17	21	19	18	3	7	13	10	8	15	6	16	5	4	12	1	20	2	11	9	14
U_H index, Eq. 4	17	21	19	18	3	10	1	20	7	16	6	8	13	2	5	4	11	12	14	9	15
D_{ML} index, Eq. 5	17	21	19	18	3	10	1	20	7	16	6	8	13	2	5	4	11	12	14	9	15
Ch index, Eq. 6	21	17	19	18	1	20	3	10	16	2	6	7	8	13	5	4	11	12	14	9	15
χ^2 index, Eq. 6	21	17	19	18	1	20	3	10	2	6	16	7	8	13	5	11	4	12	14	9	15

TABLE II

RESULTS OF RANKING METHODS ON THE ABALONE DATASET WITH THE EQUAL-WIDTH DISCRETIZATION.

Method	**Most – Least Important**								
ADC index, Eq. 1	9	4	3	6	8	7	5	2	1
U_S index, Eq. 3	5	9	4	3	8	6	2	7	1
U_H index, Eq. 4	9	5	4	6	3	8	7	2	1
D_{ML} index, Eq. 5	9	5	4	6	3	8	7	2	1
Ch index, Eq. 6	9	4	3	6	8	7	5	2	1
χ^2 index, Eq. 7	4	3	9	6	8	5	7	2	1

sufficient, producing one ranking for each index. For real application this could lead to some overfitting, therefore ranking and classification should be done separately for each training partition. Good generalization may be obtained by selecting only those features that were highly ranked in all data partitions.

The ranking of features for 6 indices is presented in Tab. 2. For the abalone dataset the quality of classification is obviously quite low, but many errors are small, getting the predicted age of the abalone wrong by one or two years. The number of data vectors for ages 6-13 years is significantly larger than for very young or old abalones, thus many larger errors are made outside this range.

Classification accuracy for the k-NN and C4.5 classifiers is presented in Fig. 2. Unfortunately U_S index selects now as the only one a rather poor feature number 5, followed by the best feature number 9. Peak accuracy is reached on the 6 best features in ranking by χ^2 index. These results are quite different than for the previous dataset. Calculations on more datasets confirm that there is no clear winner among the entropy and χ^2 indices.

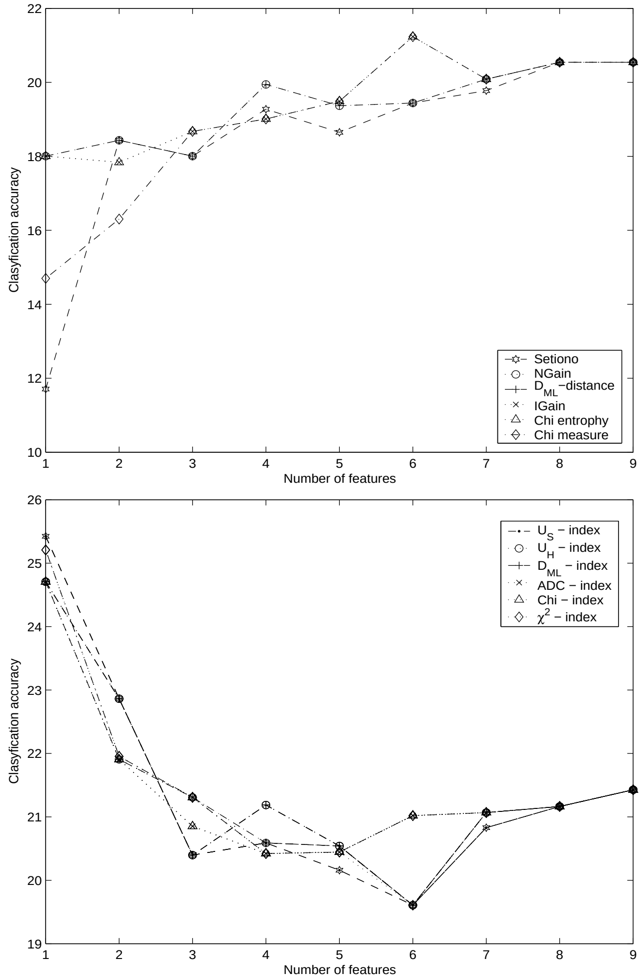

Fig. 2. Classification accuracy for the abalone dataset: upper figure – 1NN classifier, lower – C4.5 .

V. CONCLUSIONS

Ranking methods may filter features leading to reduced dimensionality of the feature space. This is especially effective for classification methods that do not have any inherent feature selections build in, such as the nearest neighbor methods or some neural networks. Five entropy-based ranking methods have been evaluated and compared with an index based on the χ^2 values. Although they all perform in a similar way (as verified on the artificial Gauss 1 and Gauss2 data), accuracy of the nearest neighbor classifier (and to some degree also the C.4.5 classifier) has been significantly influenced by the ranking index. For the two experiments presented here, and other experiments that have not been reported, different ranking methods emerge as the winner. The simple χ^2 statistical test gives similar results to the entropy based indices, reaching for the Abalone data highest accuracy. From computational point of view this index is slightly less expensive then entropy based indices, although in practice this may not be so important.

The algorithms and datasets used in this paper were selected according to precise criteria: entropy-based algorithms and several datasets either real or artificial with nominal, binary and continuous features. The two real datasets illustrated the fact that the best index (U_S) on one of them may select the worst first feature on the other. The classifiers used for evaluation of feature subsets generated by ranking procedures were deterministic, to avoid additional source of variance.

What conclusions may one draw from this study? There is no best ranking index, for different datasets and different classifiers accuracy curves as a function of the number of features used may significantly differ. Evaluation of ranking indices is fast. The only way to be sure that the highest accuracy is obtained in practical problems requires testing a given classifier on a number of feature subsets, obtained from different ranking indices. The number of tests needed to find the best feature subset is very small comparing to the cost of wrapper approach for larger number of features.

Several improvements of the ranking methods presented here are possible:

- Results of ranking algorithms depend on discretization procedure for continuous features, therefore better discretization should be used.
- Crossvalidation techniqes may be used to select features that are important in rankings on all partitions.
- Features with lowest ranking values of various indices in all crossvalidations may be safely rejected.
- The remaining features should be analyzed with selection methods that allow for elimination of redundant and correlated features.
- More ranking indices similar to χ^2 that evaluate similarity of statistical distributions may be introduced.

These conclusions and recommendations will be tested on larger datasets using various classification algorithms in the near future.

Acknowledgement: the authors wish to thank Adam Kachel for his help in implementing some ranking algorithms and making the present experiments possible.

References

[1] Breiman L., Friedman J.H., Olshen R.H., Stone C.J., *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[2] Chi J., "Entropy based feature evaluation and selection technique", Proc. of 4^{th} Australian Conf. on Neural Networks (ACNN'93), pp. 181-196, 1993.

[3] Duch W., Adamczak R., Grąbczewski K., "A new methodology of extraction, optimization and application of crisp and fuzzy logical rules." IEEE Transactions on Neural Networks, vol. 12, pp. 277-306, 2001.

[4] Duch W., Biesiada J., Winiarski T., Grudziński T., Grąbczewski K., "Feature selection based on information theory filters and feature elimination wrapper methods." In: Neural Networks and Soft Computing (eds. L. Rutkowski and J. Kacprzyk), Advances in Soft Computing, Physica Verlag (Springer), pp. 173-176, 2002.

[5] Duch W., Winiarski T., Biesiada J., Kachel A., "Feature Ranking, Selection and Discretization". Int. Conf. on Artificial Neural Networks (ICANN) and Int. Conf. on Neural Information Processing (ICONIP), Istanbul, June 2003, pp. 251-254.

[6] Hall M.A., *Correlation based feature selection for machine learning*. PhD thesis, Dept. of Comp. Science, Univ. of Waikato, Hamilton, New Zealand (1998)

[7] Kohavi R., John G.H., "Wrappers for feature Subset Selection." Artificial Intelligence, vol. 97, pp 273-324, 1997.

[8] Lorenzo M., Hernández C., Méndez J., "Atributte selections through a measurement based on information theory". 7^{th} Conferencia de la Asociación Española para la Inteligencia Artificial, (CAEPIA 1997), pp. 469-478, 1997.

[9] Lopez de Mantaras R., "A Distance-Based Attribute Selecting Measure for Decision Tree Induction", Machine Learning vol. 6, pp. 81-92, 1991.

[10] Mertz C.J., Murphy P.M., UCI repository of machine learning databases http://www.ics.uci.edu.pl/~mlearn/MLRespository.html. Irvine, CA: University of California, Department of Information and Computer Science.

[11] Setion R., Liu H., "Improving backpropagation learning with feature selection". Applied Intelligence: The International Journal of Artifical Intelligence, Neural Networks, and Complex Problem-Solving Technologies, vol. 6, pp. 129-139, 1996.

[12] Setion R., Liu H., "Chi2: Feature selection and discretization of numeric attributes". In: Proc. 7th IEEE International Conf. on Tools with Artificial Intelligence, Washington D.C., pp. 388-391, Nov. 1995.

[13] Shanonn C.E., Weaver W., *The Mathematical Theory of Communication*. Urbana, IL, University of Illinois Press, 1946.

[14] Shridhar D.V., Bartlett E.B., Seagrave R.C., "Information theoretic subset selection." Computers in Chemical Engineering, vol. 22, pp. 613-626, 1998.

[15] Quinlan J.R. *C4.5: Programs for machine learning*. San Mateo, Morgan Kaufman, 1993.