

Exploration and Reduction of the Feature Space by Hierarchical Clustering

Dino Ienco*

Rosa Meo†

Abstract

In this paper we propose and test the use of hierarchical clustering for feature selection. The clustering method is Ward's with a distance measure based on Goodman-Kruskal tau. We motivate the choice of this measure and compare it with other ones. Our hierarchical clustering is applied to over 40 data-sets from UCI archive. The proposed approach is interesting from many viewpoints. First, it produces the feature subsets dendrogram which serves as a valuable tool to study relevance relationships among features. Secondly, the dendrogram is used in a feature selection algorithm to select the best features by a wrapper method. Experiments were run with three different families of classifiers: Naive Bayes, decision trees and k nearest neighbours. Our method allows all the three classifiers to generally outperform their corresponding ones without feature selection. We compare our feature selection with other state-of-the-art methods, obtaining on average a better classification accuracy, though obtaining a lower reduction in the number of features. Moreover, differently from other approaches for feature selection, our method does not require any parameter tuning.

1 Introduction

Feature selection is a widely recognised important task in machine learning. It is well-known for instance that classification in high-dimensional data-sets improves its accuracy thanks to feature selection since the chance of overfitting increases with the number of features. In addition, tasks based on proximity computations such as distance-based clustering and k-nn classifiers observe a degradation of the quality of their results. This is due to the fact that the relative difference in distances decreases with the number of features, a problem known as the curse of dimensionality. Degradation of costs and execution times is also a relevant issue since the feature space is exponentially large in the number of features and the proximity computations are proportional to the number of features. Therefore, some features should be filtered out especially when carry redundant information for the task at hand. On the other side, predictive models based on a lower number of features

are more easy to understand and interpret by the human expert and descriptive models can be visualized better in a lower number of dimensions [5, 12].

1.1 Distinctive issues of this paper and comparison with related work. Some approaches to feature selection adopt a filter method [7, 17, 23] based on an attempt to immediately derive non redundant and relevant features for the task at hand (e.g. for prediction of the classes). Others are based instead on a wrapper approach [13, 16, 19] in which the subset of features is selected also on the basis of the accuracy that a classifier induced by a previously chosen learner is able to obtain with that subset of features. In particular, [13, 16] claim that the optimal result on the relevance of the features for the task at hand greatly depends both on the learner and the domain. Therefore, a big effort in exploration of the feature space and in understanding of the features relative importance is necessary.

In this paper, similarly to [23] we perform a selection of the features on the basis of their relevance and redundancy. Similarly to them we choose to determine feature relevance with a measure of predictive power. Differently, instead of using Symmetrical Uncertainty - a measure based on the mutual information between the features normalized by the sum of their entropy - we choose a measure of association of a feature to the class that preserves the distribution of the class [11]: it is Goodman-Kruskal τ (see Section 2.1 for details on τ). A further difference with [23] is that we use a feature clustering approach instead of a ranking one. In the experimental section we show that our approach allows in practice a better classification accuracy.

Clustering, and more in particular incremental one, has not been often adopted in feature selection but in the context of text classification. See for instance [22] for a survey and [2, 7, 8] for examples adopting clustering approaches. In text classification hierarchical clustering is a more popular method for dimensionality reduction since it tries to organize words in groups according to their similar meaning. Few works on dimensionality reduction by clustering exist in classification of structured examples. [4, 9] are respectively based on DBSCAN [10] and AGNES [14] algorithms.

[4] is a pioneer work on feature selection by hierar-

*Dipartimento di Informatica, Università di Torino, Italy.

†Dipartimento di Informatica, Università di Torino, Italy.

chical clustering. It is based on Barthélemy-Montjardet distance, which has an opposite behavior of Pearson χ^2 coefficient. Unfortunately, the cluster evaluation measure based on Barthélemy-Montjardet distance does not identify a feature subset that allows the classifiers to improve their original performance accuracy. Furthermore, even compared with other feature selection methods, the obtained accuracy is lower.

In our work, similarly to [4] we adopt a hierarchical, agglomerative clustering approach that organizes the features into progressively larger groups. In our paper, feature groups are based on the capability of some features in the group to predict some other features in the group. In this way, the features in a same cluster form a set of features from which a single representative is selected - the other features in the cluster are judged redundant since their values can be predicted. Again, representatives are chosen on the basis of their capability to predict the class. From experimental results of Section 3.4 we show that this mechanism of feature selection based on hierarchical clustering is effective: it allows to improve the classification performances of different classifiers with respect to their original performances without feature selection. In Section 3.2, we also show the classification accuracy obtainable by our feature selection method changing the cluster evaluation measure. We show that Goodman-Kruskal τ , based on the feature predictive power, allows classifiers to reach on average a higher accuracy.

A distinctive point in our work is that we apply hierarchical clustering which produces a dendrogram of feature subsets. In many fields, such as biological sciences, the dendrogram corresponds to a meaningful taxonomy. We believe this dendrogram is determinant to help the exploration and understanding of the feature space and to ease the study of the relative importance of the features. At lower levels of the dendrogram there are the clusters with the most redundant features since the most redundant features are grouped first. At lower levels, yet, a higher number of clusters is present and more representatives are kept. Higher levels instead, have a lower number of clusters and fewer representatives are chosen. As a disadvantage, clusters at higher levels might contain less redundant features. As a consequence, the choice of the cluster representatives could generate a higher loss of information. Differently from [4], we pose the problem of the determination of the right level of the dendrogram. In our work, this is done with a wrapper approach, according to the performance reached by the adopted learner in the particular domain.

Our approach is hybrid. Hierarchical clustering, followed by the choice of a representative from each cluster, works as a filter approach, while the actual selection of

the most appropriate level in the dendrogram follows a wrapper approach. We believe this solution is interesting, since clustering (guided by a measure of predictive power) solves the problem of features redundancy while the choice of the representatives determines the relevance of the features for the task at hand. On the other side, as we will see, our solution does not require any special nor complex parameter tuning process that some other competitive approaches require (see for instance, **ReliefF** [18]). Furthermore, hierarchical clustering produces the dendrogram which is a powerful conceptual tool for the exploration of the feature space.

Overall, we performed extensive experiments, both empirical and statistical, to validate the presented approach and its results.

1.2 Organization of the paper. In Section 2 we present our method based on Goodman-Kruskal τ association measure (revised in Section 2.1) and on Ward hierarchical clustering (Section 2.2). Our hybrid method is discussed in Section 2.3. Experimental results are presented in Section 3. The proposed approach is statistically validated in Section 3.1 and a comparison of results allowed by τ and other measures is presented in Section 3.2. The benefits of obtaining a feature subsets dendrogram are discussed in Section 3.3. The increase in accuracy performance of three different families of learners - Naive Bayes, Decision Trees and K-Nearest Neighbors - coupled with our feature selection as a preprocessing step, is reported in Section 3.4. Section 3.5 and Section 3.6 report a comparison with other methods, respectively filters and wrappers. According to these experiments, hierarchical clustering produces better accuracy results than approaches based only on features ranking (see for instance [23]). Finally, Section 4 draws conclusions.

2 Feature selection by hierarchical clustering

In this section we present a clustering-based method for the investigation of the relationships between features in a data-set, denoted by DB. Then we use this approach to build an hybrid feature selection algorithm.

2.1 Goodman-Kruskal τ as a cluster evaluation measure. Goodman-Kruskal τ , as other measures proposed by the same authors, can be described as a measure of proportional reduction in the error [11]. It describes the association between two categorical variables - a dependent and an independent one - in terms of the increase in the probability of correct prediction of the category of the dependent variable when information is supplied about the category of the other, independent variable. τ is intended for use with any categori-

cal variable while other measures (eg., γ) are intended for ordinal ones. In the present work we applied τ to any type of features, having performed a pre-processing (discretization step) of any non categorical, continuous feature. In future work we intend to investigate on the use of others, similar measures, for ordinal types which would avoid the necessity of the discretization step.

In order to review the meaning of τ consider the table in Figure 1 whose cells at the intersection of the row I_i with the column D_j contain the frequencies of database examples having the category I_i of the feature I and category D_j of the feature D .

	D_1	D_2	...	D_j	D_n	Total
I_1	n_{11}	n_{12}	...	n_{1j}	n_{1n}	$I_1 \text{ Total}$
I_2	n_{21}	n_{22}	...	n_{2j}	n_{2n}	$I_2 \text{ Total}$
...
I_i	n_{i1}	n_{i2}	...	n_{ij}	n_{in}	$I_i \text{ Total}$
...
I_m	n_{m1}	n_{m2}	...	n_{mj}	n_{mn}	$I_m \text{ Total}$
Total	D_1	D_2	...	D_j	D_n	Matrix
	Total	Total	...	Total	Total	Total

Figure 1: Cross-classification table of two categorical features I and D .

τ determines the predictive power of a feature I , considered as an independent variable, for the prediction of the dependent variable D . The prediction power of I is computed as a function of the error in the classification of D .

The prediction error in the classification of D is first computed when we do not have any knowledge on the variable I . E_D denotes this error here. The reduction of this error allowed by I is obtained by subtraction from E_D of the error in the classification of D that we make when we have the knowledge of the value of I in any database example. $E_D|I$ denotes this latter error. The proportional reduction in prediction error of D given I , here called $\tau_{I \rightsquigarrow D}$, is computed by:

$$(2.1) \quad \tau_{I \rightsquigarrow D} = \frac{E_D - E_D|I}{E_D}$$

E_D and $E_D|I$ are computed by a predictor which uses information from the cross-classification frequencies and tries to reduce as much as possible the prediction error. In the prediction, it also preserves the dependent variable distribution (relative frequencies of the predicted categories D_j) in the following way: when no knowledge is given on the category of I , category D_j is predicted with the relative frequency determined by $\frac{D_j \text{ Total}}{\text{Total}}$; otherwise, when it is known that I_i is the category of variable I , category D_j is predicted with the

relative frequency determined by $\frac{n_{ij}}{I_i \text{ Total}}$. Therefore, error E_D is determined by:

$$(2.2) \quad E_D = \sum_j \left(\frac{\text{Total} - D_j \text{ Total}}{\text{Total}} \cdot D_j \text{ Total} \right)$$

while error $E_D|I$ is:

$$(2.3) \quad E_D|I = \sum_i \sum_j \left(\frac{I_i \text{ Total} - n_{ij}}{I_i \text{ Total}} \cdot n_{ij} \right)$$

Analyzing the properties of τ , we can see that it satisfies many desirable properties for a measure of association. For instance, it is invariant by rows and columns permutation. Secondly, it seems a good measure for the association between two features since it takes values between (0,1) and it is 0 if and *only if* there is independence between the features (solves the problem of zero uniqueness). Furthermore, it has the benefit that it can be interpreted since it has an operational meaning - it corresponds to the relative reduction in the prediction error in a way that preserves the class distribution.

However it is easy to notice that it is not symmetrical. It is $\tau_{I \rightsquigarrow D} \neq \tau_{D \rightsquigarrow I}$. This could be a problem for the adoption of τ as measure of proximity between features to be later used in distance calculations. We solve the problem with the following definition.

We use a function of τ as measure of distance *dist*.

DEFINITION 2.1. We call $d_{I,D} = 1 - \tau_{I \rightsquigarrow D}$. We know that $\tau_{I \rightsquigarrow D} \neq \tau_{D \rightsquigarrow I}$. Then, the distance between two features I, D is defined by $\text{dist}(I, D) = \max\{d_{I,D}, d_{D,I}\}$.

Note that the domain of *dist* is still in (0,1) but it increases as the association between the two features decreases. Furthermore, it is $\text{dist}(D, D) = 0$ for any D . Therefore, it satisfies the properties of positive semi-definiteness. In addition, with Definition 2.1, it satisfies also the symmetrical property.

We will see in Section 3.2 that τ , used in conjunction with hierarchical clustering as measure of feature relevance, allows classifiers to reach higher accuracies than other measures such as Pearson χ^2 coefficient and SU (Symmetrical Uncertainty) - based on measures of information theory.

2.2 Ward's hierarchical clustering method

Ward's hierarchical method is one of the hierarchical clustering methods most used in literature [1]. It is a greedy, agglomerative hierarchical method, that determines a diagram - the dendrogram - that records the sequence of fusions of clusters into larger clusters. At

the end of the process, a unique cluster - with all the population - is usually determined ¹. See Figure 5 for an example of a dendrogram produced by a hierarchical, agglomerative clustering applied to the population of the features in a particular data-set, Mushroom [3]. Note that in Figure 5 numbers identify elements (single features) in the population of the features whose meaning is reported in Figure 4.

Ward's method is iterative. An objective function determines the best candidate clusters that will be merged in a new cluster, at each iteration of the algorithm (corresponding to a dendrogram level). The objective function is based on the computation of an overall, global measure of goodness of each candidate clustering solution. In turn, the global measure of each solution is given by a summation, over all clusters of the solution, of a measure of cluster cohesion. Cluster cohesion could be computed in many ways. One, well-known measure, is the sum of squared errors (SSE) within each cluster or differently viewed, the variance of the cluster elements (let call it W_i for cluster i). When two clusters are merged, in the agglomerative hierarchical clustering, the overall global measure increases. All the candidate solutions are computed but the best one is determined by the minimum increase in the objective function. Ward's algorithm, like other clustering hierarchical algorithms, takes in input a matrix containing the distances (**dist**) between any pair of elements (in our case the database features).

At the core of the method lies the computation of the cluster cohesion of each new cluster and the update of the matrix of distances for the inclusion of the new cluster. The new cluster cohesion is computed on the basis of the cohesions of the clusters that are considered for the merge, as follows

$$(2.4) W_{ir} = \frac{(|i| + |p|)W_{ip} + (|i| + |q|)W_{iq} - |i|W_{pq}}{|i| + |r|}$$

where p and q represent two clusters, r represents the new cluster formed by the fusion of cluster p and q , while i represents any other cluster. Notation $|i|$ is for the cardinality (number of elements) of cluster i while W_{ij} denotes the cohesion of the cluster that would be obtained if cluster i and j were merged. [20] shows that equation 2.4 can be easily generalized to embed in the algorithm any criteria for cluster fusion, such as, for instance, Nearest Neighbor (also known as MIN or SingleLink), Furthest Neighbor (also MAX or CompleteLink), Median, Group Average and Centroid.

¹The final number of clusters could be also any $k < N$ (N =total number of elements), if a fixed number k of clusters is desired.

2.3 An hybrid feature selection approach. We have seen in previous sections Ward's method for hierarchical clustering and the Goodman-Kruskal τ association measure.

We introduce here the notation for this paper.

- DB: the set of objects in the database;
- F: the set of features (or attributes) of DB;
- I,D,R,...: capital letters denote any single feature in F;
- M: the matrix of distances between features computed by measure *dist* of Definition 2.1;
- C_i : a feature subset of F; we call it also a cluster of features;
- T: a dendrogram (or tree diagram) that organizes the feature set F in subsets and represents the arrangement of the clusters produced by a hierarchical clustering algorithm;
- DB[C_i]: the database projection over the features subset C_i .

The following definition states the properties of the dendrogram.

DEFINITION 2.2. *The dendrogram T is a set of cluster pairs (C_i, C_j) (in a father-child relationship) where C_i and C_j belong to the power-set of F and such that they satisfy the following properties:*

1. $C_j \subset C_i$.
2. taken two cluster pairs P_1 and P_2 in T, such that $P_1=(C_i, C_j)$ and $P_2=(C_i, C_k)$ the following statements hold: $C_j \cap C_k = \emptyset$ and $C_i = C_j \cup C_k$.

The dendrogram can be further organized in levels. Each level l contains a set \mathcal{L}_l of clusters C_i that satisfy the following properties:

1. $\bigcup_{C_i \in \mathcal{L}_l} C_i = F$.
2. $\forall C_j, C_k \in \mathcal{L}_l : C_j \cap C_k = \emptyset$.

In our approach we combine hierarchical clustering and the Goodman-Kruskal τ association measure to investigate the feature space of a data-set. With hierarchical clustering we build the feature dendrogram that groups in each cluster the most correlated features (algorithm 1). This feature dendrogram can be used to study the particular characteristics of the feature set. It is a very useful and intuitive tool that allows us to understand the feature relationships by a taxonomic organization: according to the chosen feature proximity measure, the most relevant features are placed at the lowest

levels while the less associated ones are added by last to existing clusters at higher levels. An expert can view the feature space from different viewpoints, through the dendrogram: she/he can analyze this structural representation and choose from each cluster, at certain level of the tree, the features judged more important. Alternatively, an heuristic procedure can be carried out to determine automatically the most relevant features of each cluster.

We use our investigation approach as an hybrid feature selection algorithm:

1. Cluster the feature space F of DB with hierarchical clustering which produces a dendrogram T ;
2. At each level l of the dendrogram T we extract the corresponding clusters $C_i \in L_l$;
3. For each cluster C_i we choose a representative R that has the higher association with the class, $\tau_{R \rightsquigarrow class}$;
4. For each level l we retain only the representative features R of each cluster C_i . Call this subset F_l ;
5. We evaluate F_l in a Wrapper method, i.e. testing the accuracy of a single classifier (e.g., Naive Bayes, decision tree, etc ..) on $DB[F_l]$. Let this accuracy be $Acc(classifier, DB[F_l])$;
6. The best feature subset $optF$ is that one that maximizes the accuracy of the given classifier, i.e. $optF(classifier, DB) = \arg \max_{F_l} Acc(classifier, DB[F_l])$.

Call $optL(classifier, DB)$ the corresponding level in the dendrogram.

The pseudo-code of the algorithm is reported as Algorithm 2. Loop starting at line 3 at first returns the clusters at a single level of the dendrogram (by procedure `getLevelPartition`) and then discards from them redundant features.

This is a solution that performs the feature selection in the complete set of features, even though at any iteration it focuses at a different course of granularity. In fact, we have seen from the properties of the dendrogram stated in Definition 2.2 that the clusters of each level are disjoint and they cover the overall set of features. This is the filtering step: it selects (and puts in `tempFSet`) the most relevant features by procedure `getRepresenter`. These ones are the representatives for the clusters in a particular level of the tree.

The wrapper step consists in the subsequent determination of the best classification accuracy, of the relative set of features and of the dendrogram level from

which the features have been selected as representatives. Its core is in procedure `Acc`: it returns the classification accuracy of the classifier (given as first argument) which is applied to the data-set (second argument) projected on the set of representative features.

Algorithm 1 HCL-Ward(M)

This procedure takes as argument the distance matrix M .

It computes a hierarchical clustering by Ward's method.

return dendrogram

Algorithm 2 HCL-FS(dendrogram, classifier, dataSet DB)

```

1: bestFeatureSet  $\leftarrow \emptyset$ 
2: best_acc  $\leftarrow 0$ 
3: for all levels  $l$  of dendrogram do
4:   Clusters[]  $\leftarrow$  getLevelPartition(dendrogram,l)
5:   tempFSet  $\leftarrow \emptyset$ 
6:   for all  $C | C \in$  Clusters[] do
7:     tempFSet  $\leftarrow$  tempFSet  $\cup$  getRepresenter(C)
8:   end for
9:   temp_acc  $\leftarrow$  Acc(classifier, dataSet DB[tempFSet])
10:  if temp_acc > best_acc then
11:    best_acc  $\leftarrow$  temp_acc
12:    optF  $\leftarrow$  tempSet
13:    optL  $\leftarrow$  l
14:  end if
15: end for
16: return optF, optL, best_acc

```

Algorithm 3 getLevelPartition(dendrogram, l)

This procedure takes as arguments:

a dendrogram and a level number.

It returns the clusters of the dendrogram at this level.

return Clusters[]

Algorithm 4 getRepresenter(cluster C)

This procedure takes as argument a cluster.

It returns the representative feature that has the maximum $\tau_{R \rightsquigarrow dataSet Class}$.

return R

3 Experimental evaluation

In this section we show some of the experimental results we have obtained by application of the proposed method of feature selection. In the experiments we adopted several data-sets from the Machine Learning archive of University of California Irvine (see [3]).

3.1 Statistical Validation of Clustering. In order to validate the hierarchical clustering algorithm we use a randomized test. For each original data-set we generate one thousand random data-sets having the same structure of the original data-set. The following method is used.

1. Randomized data-set generation:
 - (a) For each instance in the data-set, we perform step 1b
 - (b) For each feature, we swap the feature value in the instance with the corresponding value in another instance, chosen randomly.
2. On each randomized data-set thus generated, we apply hierarchical clustering on the basis of proximity matrix M based on distance function defined in Definition 2.1.
3. For each original data-set we determine optL by application of our algorithm of feature selection HCL-FS. Then we calculate the average Within cluster sum of squared errors W_k (an opposite measure to the cluster cohesion), over all the clusters at level optL . Let this value be μ . This is an overall measure of the clustering solution.
4. Similarly, for each randomized data-set, we compute the average W_k on the clusters at level optL . Let be X the random variable assuming the value of W_k computed on each data-set.
5. We compute the mean \bar{X} of X and the standard deviation over all the randomized data-sets.
6. We apply the t-test for the mean with the null hypothesis H_0 that $\mu = \bar{X}$.

Notice that with the randomization at step 1 we obtain that the features of the randomized data-sets have the same cardinality of the corresponding features in the original data-sets.

In the table of Figure 2 we report the results. We can see that we reject the null hypothesis H_0 , at significance level of 0.01%, in all the cases. We can see that the statistic value is very high (corresponding to p -values always lower than 10^{-9}). This means that our clustering method does not cluster features that are not associated.

In our experiment we have seen that, in absolute value, the minimum W_k in the randomized data-sets is always higher than any W_k in the original data-set. This means that cluster cohesion in randomized data is effectively lower than in real data - a fact that validates the determined clusters.

<i>Data-Set</i>	<i>Value of t-Test</i>
anneal	5806.53230
audiology	100.72160
autos	7184.86850
breast-cancer	583.77470
colic	2275.09820
contact-lenses	-112.69340
credit-a	30987.18833
credit-g	1019.41450
heart-c	660.06510
hepatitis	506.53860
HO	2229.63450
labor	453.42550
lymph	397.60521
mushroom	96135.58690
zoo	1503.79395

Figure 2: t-test (df: 999) for clustering evaluation on randomized data-sets.

3.2 Comparisons of τ with other proximity measures. In Figure 3 we show the results of other experiments whose aim is to validate empirically the choice of τ as a cluster evaluation measure. We have run hierarchical clustering to 42 data-sets from UCI Archive but each time we have changed the proximity measure. We want to do a comparison between τ and a set of measures frequently adopted in literature. They are:

- Goodman and Kruskal's τ (the measure used in our work);
- Pearson Chi Square test;
- Symmetric Uncertainty (SU).

The Pearson Chi Square test finds the difference between the observed frequency of occurrence of each value of two variables and the theoretical one, under the hypothesis that the two variables are independent [15].

$$(3.5) \quad \chi^2_{2n-1} = \sum_{i=1}^m \sum_{j=1}^n \frac{(n_{ij} - \frac{D_j \text{Total} * I_i \text{Total}}{\text{MatrixTotal}})^2}{\frac{D_j \text{Total} * I_i \text{Total}}{\text{MatrixTotal}}}$$

The Symmetric Uncertainty is a measure derived from information theory. It is the Information Gain normalized by the sum of entropy of the two variables.

$$(3.6) \quad SU_{ID} = \frac{IG(I, D)}{H(I) + H(D)}$$

where $H(I)$ and $H(D)$ are the entropies of I and D , respectively, defined by

$$(3.7) \quad H(X) = \sum_{x \in X} P(x) \cdot \log\left(\frac{1}{P(x)}\right)$$

We observe that τ , coupled with HCL FS, is more stable than the other two measures. In particular, the use of τ , as proximity measure, for Naive Bayes and J48 obtains better results than the other two measures, even though with not statistically significant improvements. Instead, for K-NN the measure that allows the best performance is Pearson χ^2 .

	<i>HCL NB</i>	<i>HCL J48</i>	<i>HCL KNN</i>
τ	81,34%	84,54%	84,28%
χ^2	81,02%	84,52%	84,31%
SU	81,16%	84,32%	84,27%

Figure 3: Mean accuracy allowed by different proximity measures.

3.3 Inspection of the Dendrogram. The Mushroom data set of UCI Archive [3] includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family. Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended (this latter class has been combined with the poisonous one). Any guide on mushrooms clearly states that there is no simple rule for the determination of the edible mushrooms. In Figure 4 we report a list of the features of this data-set, each identified by a number.

We applied HCL-Ward algorithm to the features of this data-set and we can see in Figure 5 the resulting dendrogram. We can understand, through this dendrogram the relationships between the features. For example we can see that feature 8 and feature 9 (respectively gill size and gill color) are immediately grouped together. This is logic because both refer to the same object, the gill of a mushroom. This suggests that at least one of them is redundant. Another similar example is the group of features 14 and 15 (respectively the stalk color of above and below the mushroom ring). Similar reasoning could apply to the features of the population and habitat (respectively features number 21 and 22) that have been clustered together by the algorithm at the first iterations. Again, this seems a meaningful fact, that corresponds to the knowledge on the particular domain.

<i>Feature Number</i>	<i>Feature name</i>
1	cap-shape
2	cap-surface
3	cap-color
4	bruises
5	odor
6	gill-attachment
7	gill-spacing
8	gill-size
9	gill-color
10	stalk-shape
11	stalk-root
12	stalk-surface-above-ring
13	stalk-surface-below-ring
14	stalk-color-above-ring
15	stalk-color-below-ring
16	veil-type
17	veil-color
18	ring-number
19	ring-type
20	spore-print-color
21	population
22	habitat

Figure 4: Features of the Mushroom data-set.

On the contrary, the dendrogram suggests that veil color and type are not so much related, according to their predictive capability, as one would tend to expect. In conclusion, the dendrogram can help a user to:

- analyze obvious relationships among features;
- discover insights on features relationships that are difficult to understand;
- build classifiers in a semi-supervised manner;
- allow analysts a certain degree of choice in the selection of the features.

In this manner the analyst can understand and investigate on the relationships between features.

3.4 Benefits of the Feature Selection. Our method of feature selection (here denoted by HCL), proposed as a pre-processing step of any classifier, though it does not allow always a strong reduction of the total number of features, always increases (or at least equals) the accuracy of the original classifiers.

In Figure 6 we show the average of classification accuracy in experiments on 42 data-sets of UCI Archive (results are shown in detail in Figure 7). We show (column **Orig Mean Acc**) the classification accuracy of the three families of classifiers: Naive Bayes, J48 decision trees and KNN - all available in Weka [21]. We compare them with the accuracy achievable by the same classifiers but coupled with our feature selection

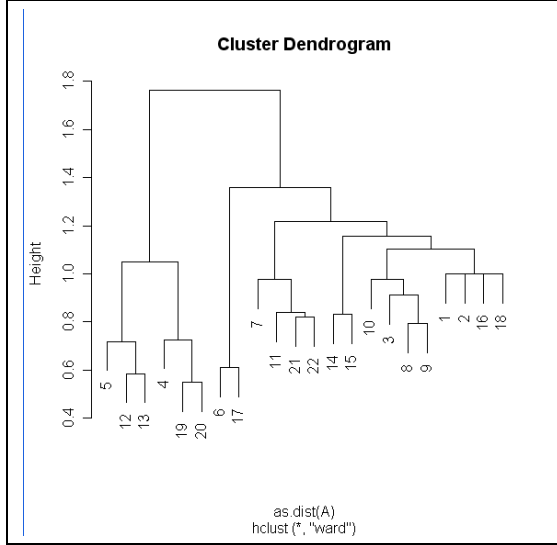


Figure 5: Dendrogram of the features of data-set Mushroom.

	<i>Orig. Mean Acc.</i>	<i>HCL Acc.</i>
<i>Naive Bayes</i>	78.33%	81.34%
<i>J48</i>	83.08%	84.54%
<i>KNN7</i>	82.51%	84.28%

Figure 6: Mean Accuracy with our FS of three classifiers: Naive Bayes, J48 decision Tree and K nearest neighbours ($k = 7$).

as a pre-processing step (column denoted by *HCL Acc.*). With all classifiers there is an enhancement of the accuracy performance. The best case is obtained with Naive Bayes (in which we obtain an improvement of 3.01%), but also in the other two cases we have a good enhancement.

3.5 Comparison with Filters. We also compare our hybrid method with two feature selection filter methods, based on Ranking. The methods are again implemented in *Weka* [21] (for which we used the default parameters setting). The two filter methods are *FCBF* (Fast Correlation-Base Feature Selection) [23] and *ReliefF* [18], two of the most adopted methods of feature selection in literature. *FCBF* is based on an heuristic search in the space of possible feature subsets. It uses Symmetrical Uncertainty to rank the features and discard redundant ones. *ReliefF* is also based on heuristic search: it uses a distance measure based on Instance based learning to choose the best subsets of features. We compare *HCL* with *ReliefF* and *FCBF* under two viewpoints: (i) the accuracy obtained by the classifiers and (ii) the number of retained features. The table

in Figure 7 shows the original accuracy of two classifiers, Naive Bayes and J48, on the 42 UCI data-sets (in columns denoted respectively by *Orig. NB* and *Orig. J48*). These accuracies are compared with the accuracy achievable by the same classifiers on the same data-sets after the feature selection by the three methods: *HCL*, *FCBF* and *ReliefF*. We can see that almost always *HCL* outperforms the other methods and always outperforms (or at least equals) original accuracy. On the contrary, we can observe that the other methods sometimes obtain a lower accuracy than the original classifier (denoted in the table of Figure 7 by the $<$ symbol). Analogous results have been obtained considering other learners. These results are statistically significant. By t-test on the paired observations the statistic value is higher than the critical t value at a significance level of $5 \cdot 10^{-3}$.

	<i>Avg Orig</i>	<i>Avg HCL NB</i>	<i>Avg HCL J48</i>
Num Feat	23.12	15.07	14.45
Mean % Red		34.81%	37.49%
		<i>Avg FCBF</i>	<i>Avg ReliefF</i>
Num Feat		6.52	16.33
Mean % Red		71.78%	29.35%

Figure 8: Comparison on Feature Reduction

The table in Figure 8 shows instead the average reduction in the number of features achieved by the three feature selection methods applied to the two classifiers. *HCL* achieves a lower average reduction but - notice - it never occurred at the expenses of a degradation of the accuracy results. On the contrary, the other methods obtained an increased reduction in the number of features but at the expenses of a degradation of the classification accuracy w.r.t. the original classifier. We believe this is an important issue to be considered for a successful feature selection method.

We also performed additional experiments between filters and our feature selection method, in order to compare classification accuracies obtainable by the two feature selection methods at the same number of features. In this case, filters win (on 2/3 of the datasets). However, we need to observe that comparisons were made at the optimal working point conditions of the best filter method (*FCBF*) and not of our method (which usually needs a higher number of features, as we have seen, in order to reach the best accuracies). We also plan to perform comparisons between them in the opposite situation, but this needs more work aimed at forcing *FCBF* to retain the same number of features as we do.

As a conclusion, we can claim that our hybrid method, composed of the following steps:

1. hierarchical clustering, viewed as a search strategy in the feature space and for identification of redun-

dant features,

2. followed by a selection of the relevant features by choice of cluster representatives

outperforms the ranking methods for feature selection - at least regarding the classification accuracy viewpoint.

3.6 Comparison with Wrappers. We performed also a comparison between our HCL method and other methods of feature selection that follow a wrapper approach. We have used the family of Weka wrappers and have used it in the following way. They search in a forward method the space of feature subsets by greedy hill-climbing. Search is augmented by a backtracking facility that allows a certain number (set equal to 5) of consecutive non-improving nodes in the greedy search. This search strategy seems to us quite advanced and powerful.

In the table of Figure 9 we report classification accuracy of the two classifiers on the 42 data-sets. Accuracies are comparable: with our method (HCL NB) it is lower than with Weka wrapper coupled with Naive Bayes, whereas our method coupled with decision trees (HCL J48) is a bit higher than Weka wrapper with J48. The heuristic search adopted by Ward method is clearly less precise than the one implemented in Weka wrapper: Ward agglomerative hierarchical method is greedy and could be trapped in local minima. We believe that the answer to the slightly inferior performances reached by our technique could be in the simpler, greedy search method we have adopted. For the future, we plan to enrich Ward hierarchical method with some backtracking facility, such as the one presented in [6].

HCL NB	HCL J48	Wrapper NB	Wrapper J48
81,34%	84,54%	82,13%	84,39%

Figure 9: Average accuracy: Comparison between our approach and Wrapper methods.

4 Conclusions

The presented approach to feature selection is an heuristic approach - lead by τ association measure. Selection of the candidate features follows a greedy, best-first, forward method - Ward hierarchical clustering. However, the feature selection itself is an hybrid approach. Clusters help to identify redundant features. From clusters, the most relevant features are chosen as representatives - a task that consists in a filter approach. The actual selection of the most appropriate combination of representatives (which corresponds to the selection of the dendrogram level) follows a wrapper approach. One of

the advantages of this solution is that it does not require any special nor complex parameter tuning process that other approaches like **ReliefF** require. Furthermore, the dendrogram produced by hierarchical clustering constitutes a useful, immediate and semantic-rich conceptual organization of the feature space. It helps the experts to explore and understand the feature space of a new problem domain.

From the extensive experiments performed - on several data-sets and learners - we have seen that our hybrid method of feature selection outperforms the ranking methods - at least regarding the classification accuracy viewpoint. In particular, we have shown that the classification accuracy achieved by different learners on many data-sets after the application of our feature selection always increases (or at least equals) the original accuracy of the same learners without feature selection. We claim that the classification accuracy achievable by the learners after the feature selection step should not be degraded as a consequence of the feature selection. This issue in the evaluation of any feature selection method is an important one and should always be considered carefully.

References

- [1] Michael R. Anderberg. *Cluster analysis for applications*. Academic, second edition, 1973.
- [2] L. Douglas Baker and Andrew Kachites McCallum. Distributional clustering of words for text classification. In *SIGIR '98: Proc. of the 21st annual Int. ACM SIGIR Conf. on Research and development in information retrieval*, pages 96–103, 1998.
- [3] C. L. Blake and C. J. Merz. UCI repository of machine learning databases, University of California, Irvine, <http://www.ics.uci.edu/~mllearn/mlrepository.html>, 1998.
- [4] Richard Butterworth, Gregory Piatetsky-Shapiro, and Dan A. Simovici. On feature selection through clustering. In *ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 581–584. IEEE Computer Society, 2005.
- [5] M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis*, 1(1-4):131–156, 1997.
- [6] Inderjit S. Dhillon, Yuqiang Guan, and J. Kogan. Iterative clustering of high dimensional text data augmented by local search. In *Proc. of IEEE Int. Conf. on Data Mining (ICDM'02)*, page 131, 2002.
- [7] Inderjit S. Dhillon, Subramanyam Mallela, and Rahul Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, 3:1265–1287, 2003.
- [8] Inderjit S. Dhillon and Dharmendra S. Modha. Concept decompositions for large sparse text data using clustering. *Mach. Learn.*, 42(1-2):143–175, 2001.

- [9] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Michael Wimmer, and Xiaowei Xu. Incremental clustering for mining in a data warehousing environment. In *Proc. of Int. Conf. on Very Large Data Bases*, 1998.
- [10] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of KDD Conference*, pages 226–231, 1996.
- [11] Leo A. Goodman and William H. Kruskal. Measures of association for cross classifications. *Journal American Statistical Association*, 49(268):732–64, December 1954.
- [12] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, 2003.
- [13] Jinjie Huang, Yunze Cai, and Xiaoming Xu. A wrapper for feature selection based on mutual information. In *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, pages 618–621. IEEE Computer Society, 2006.
- [14] L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 1990.
- [15] Maurice Kendall, Alan Stuart, and J. Keith Ord. *Kendall's Advanced Theory of Statistics*. Oxford University Press, 1994.
- [16] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.
- [17] I. Kononenko and M. Robnik-Sikonja. Relief for estimation and discretization of attributes in classification, regression and ilp problems, 1996.
- [18] Marko Robnik-Sikonja and Igor Kononenko. Theoretical and empirical analysis of relief and rrelief. *Machine Learning*, 53:23–69, 2003.
- [19] Juan Torres, Ashraf Saad, and Elliot Moore. *Application of a GA/Bayesian filter-wrapper feature selection method to classification of clinical depression from speech data*, pages 115–121. Springer-Verlag, 2007.
- [20] David Wishart. Note: An algorithm for hierarchical classifications. *Biometrics Journal*, 25(1):165–170, March 1969.
- [21] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, second edition, 2005.
- [22] Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *ICML '97: Proc. of the Fourteenth International Conference on Machine Learning*, pages 412–420, 1997.
- [23] Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *J. Mach. Learn. Res.*, 5:1205–1224, 2004.

<i>Data-set</i>	<i>Orig. NB</i>	<i>HCL NB</i>	<i>FCBF NB</i>	<i>ReliefF NB</i>	<i>Orig. J48</i>	<i>HCL J48</i>	<i>FCBF J48</i>	<i>ReliefF J48</i>
anneal	86.30%	90.31%	86.97%	86.19%<	98.44%	98.44%	96.88%	98.44%
audiology	73.45%	73.89%	73.89%	73.89%	77.88%	78.32%	77.88%	77.88%
autos	56.10%	63.42%	53.17%<	56.10%	81.95%	81.95%	69.76%<	81.95%
badges2	99.66%	100.00%	100.00%	100%	100.00%	100.00%	100.00%	100.00%
balance-scale	90.40%	90.40%	90.40%	90.40%	76.64%	76.64%	76.64%	76.64%
breast-cancer	71.68%	74.48%	70.28%<	73.08%	75.52%	75.87%	69.93%<	75.52%
breast-w	95.99%	95.99%	95.85%<	95.99%	94.56%	95.85%	95.85%	94.56%
colic	77.99%	84.51%	83.15%	79.89%	85.33%	85.87%	81.79%<	85.33%
contact-lenses	70.83%	87.50%	70.83%	83.33%	83.33%	87.50%	83.33%	83.33%
credit-a	77.68%	86.09%	75.22%<	86.38%	86.09%	87.54%	85.07%<	86.23%
credit-g	75.40%	75.60%	74.40%<	75.00%<	70.50%	70.50%	70.50%	72.20%
diabetes	76.30%	77.47%	77.47%	76.43%	73.83%	75.26%	74.87%	74.48%
glass	48.60%	53.27%	42.52%<	48.60%	66.82%	67.76%	69.16%	66.82%
heart-c	83.50%	83.83%	84.49%	83.83%	77.56%	80.53%	77.23%<	76.57%<
hepatitis	84.52%	86.45%	84.52%	84.52%	83.87%	83.87%	82.58%<	81.29%<
HO	77.99%	84.51%	83.15%	79.89%	85.33%	85.87%	81.79%<	85.33%
hypothyroid	95.28%	95.28%	94.62%<	95.10%	99.58%	99.58%	97.32%<	99.34%<
ionosphere	82.62%	86.61%	88.89%	82.62%	91.45%	91.45%	90.03%<	91.17%<
iris	96.00%	96.00%	96.00%	96.00%	96.00%	96.00%	96.00%	96.00%
kropt	36.01%	36.01%	33.26%<	36.01%	56.58%	56.58%	52.22%<	56.58%
kr-vs-kp	87.89%	91.52%	91.99%	88.77%	99.44%	99.47%	94.06%<	97.78%<
labor	89.47%	94.74%	89.47%	91.23%	73.68%	85.97%	77.19%	75.44%
letter	64.12%	65.09%	65.52%	64.12%	87.98%	88.22%	88.47%	88.14%
lymph	83.11%	83.78%	79.73%<	81.08%<	77.03%	79.73%	75.68%<	77.03%
monk2	56.81%	62.13%	59.17%	61.54%	56.21%	62.13%	62.13%	57.99%
mushroom	95.83%	98.62%	98.52%	95.83%	100.00%	100.00%	99.02%<	100.00%
primary-tumor	50.15%	50.15%	47.20%<	49.85%<	39.82%	44.25%	41.89%	41.00%
segment	80.22%	81.91%	86.67%	83.25%	96.93%	97.06%	97.01%	96.97%
sick	92.60%	96.55%	96.53%	94.43%	98.81%	98.83%	97.45%<	97.61%<
sonar	67.79%	73.56%	70.67%	71.15%	71.15%	76.44%	75.48%	72.60%
soybean	92.97%	92.97%	89.60%<	92.97%	91.50%	92.24%	91.07%<	91.51%
spambase	79.29%	79.79%	76.87%<	67.83%<	92.98%	92.98%	93.31%	76.16%<
splice	95.30%	95.58%	96.14%	95.36%	94.08%	94.11%	94.51%	94.39%
ticdata-categ	85.52%	94.02%	93.90%	80.83%<	93.97%	94.02%	94.02%	94.02%
titanic	77.87%	77.87%	77.60%<	67.92%<	78.92%	78.92%	77.60%<	67.51%<
vehicle	44.80%	53.19%	43.26%<	44.80%	72.46%	74.70%	58.27%<	72.70%
vote	90.12%	95.63%	96.09%	90.11%<	96.32%	96.32%	96.09%<	96.32%
vowel	63.74%	67.78%	63.43%<	64.65%	81.52%	81.52%	81.11%<	80.10%<
waveform-5000	80.00%	80.84%	77.90%<	79.90%<	75.08%	75.46%	76.88%	76.66%
weather	64.29%	64.29%	57.14%<	57.14%<	64.29%	71.43%	42.86%<	71.43%
wine	96.63%	97.75%	97.75%	96.63%	93.82%	95.51%	93.82%	93.82%
zoo	95.05%	97.03%	93.07%<	95.05%	92.08%	96.04%	92.08%	92.08%
Mean	78.33%	81.34%	78.75%	78.52%	83.08%	84.54%	81.64%	82.64%
Mean Incr. Acc		3.01%	0.42%	0.19%		1.46%	-1.44%	-0.44%

Figure 7: Comparing the accuracy of Naive Bayes and J48 decision trees coupled with our method (HCL FS) vs. FCBF and ReliefF