On Feature Selection: a New Filter Model

Marc Sebban
Department of Juridical and Economical Sciences
West Indies and Guiana University
Campus de Fouillole 97159 Pointe à Pitre (FRANCE)
msebban@univ-ag.fr

Abstract

We focus on the filter approach of feature selection. We exploit geometrical information of the learning set to build an estimation criterion based on a quadratic entropy. The distribution of this criterion is approximately normal, that allows the construction of a non parametrical statistical test to assess the relevance of feature subsets. We use the critical threshold of this test, called the test of *Relative Certainty Gain*, in a forward selection algorithm. We present some experimental results both on synthetic and natural domains of the UCI database repository, which show significantly improvements on the accuracy estimates.

Introduction

While the problem of feature selection has always been at the center of statistic researches, it is only recently that this problem received attention in computer science. Beyond the intention of improving the performance of their algorithms, machine learning researchers studied feature selection methods to face the explosion of data (not always relevant) provided by recent data collecting technologies (the Web for instance).

From a theoretical standpoint, the selection of a good feature subset is of little interest. Actually, a Bayesian classifier is monotonic, *i.e.*, adding features can not decrease the model's performance. This is generally true only for infinite learning sets for which the estimate errors can be ignored. In fact, practical algorithms not always being perfect, the monotonicity assumption rarely holds (Kohavi 1994). Thus, irrelevant or weakly relevant features may reduce the accuracy of the model. A study in (Thrun et al. 1991) shows that with the C4.5 algorithm (Quinlan 1993) non deletion of weakly relevant features generates deeper decision trees with lower performances than those obtained without these features. In (Aha 1992), the author shows that the storage of the IB3 algorithm increases exponentially with the number of irrelevant features. Same sort of conclusions are presented in (Langley and Iba 1993). These results have encouraged scientists to elaborate sophisticated feature selection methods allowing to:

- Reduce classifier's cost and complexity.
- Improve model accuracy.
- Improve the visualization and comprehensibility of induced concepts.

According to the terminology proposed in (John, Kohavi and Pfleger 1994), two approaches are available: the *wrapper* and *filter* models.

In filter models, the accuracy of the future induced classifier is assessed using statistical techniques. The method "filter out" irrelevant features *before the induction process*. In wrapper methods, we search for a good subset of features *using the induction algorithm*. The principle is generally based on the optimization of the accuracy rate, estimated by one of the following methods: *holdout*, *cross-validation* (Kohavi 1995), or *bootstrap* (Efron and Tibshirani 1993).

Whatever the method of feature selection we use, the goal is always to assess the *relevance* of alternative subsets. A survey of relevance definitions is proposed in (Blum and Langley 1997).

In this article, we consider the filter approach to find relevant features. We will explain in detail arguments about this choice. We exploit characteristics of a neighborhood graph built on the learning set, to compute a new estimation criterion based on a quadratic entropy. We show that the distribution of this criterion is approximately normal, that allows the construction of a non parametrical test to assess the quality of feature subsets. We use this statistical test (more precisely the critical threshold) in a forward selection. Finally, we present some experimental results on benchmarks of the UCI database repository, comparing performances of selected feature subsets with results obtained in the original spaces.

Feature Selection and Filter Model

Presentation

Given a p-dimensional representation space, where p is the number of features characterizing a S set of n learning instances. Each instance ω_i is represented by a p-dimensional input vector $X(\omega_i) = (x_{i1}, x_{i2}, .., x_{ip})$, and by a label $Y(\omega_i) \in Y, Y = \{y_1, y_2, ..., y_k\}$. We would like to build a hypothesis h so that $h(x_{i1}, x_{i2}, .., x_{ip}) = Y(\omega_i)$ the more often. Even if h can be built from all the attributes, we would like the selected hypothesis to use only a small subset of features for reasons mentioned above.

The feature selection problem in general is NP-hard. Actually, there are 2^p different combinations to test. Then, the optimal selection can only be done with few features. In front of large representation spaces, a lot of works in machine learning has developed a large number of heuristics for performing the search of a "good subset" efficiently. According to the paper of Langley (Langley 94), four basic issues determine the nature of the heuristic search process:

- The *starting point* in the space: with an empty space (forward selection) or with all the features (backward selection).
- The *organization of the search*: addition or deletion of an attribute at each stage, never reconsidering the previous choice.
- The *strategy used to evaluate* alternative subsets of attributes (filter or wrapper model).
- The *criterion for halting search* through the space of feature subsets. Probably the simplest solution consists in fixing in advance the feature subset size k. We propose in our approach a more sophisticated statistical criterion.

We consider in this article a filter approach with a statistical criterion for halting search. We explain this position by the following arguments: while the wrapper approaches often provide better accuracy estimates than a statistical measure, they tend to be more computationally expensive. Moreover, the bias of the learning algorithm does interact with the bias inherent in the feature selection algorithm. In fact, we think that the relevance of a feature subset must not depend on a given classifier. This relevance is an intrinsic property of the concept represented by the attributes. That's why we believe (it is a postulate !) that the feature selection must be an *a priori* preprocessing, and requires then a filter approach. This preprocessing step must use general characteristics of the training set. It requires statistical or information criteria to measure the feature relevance.

Estimation Criteria

We present here different estimation criteria used for feature selection or feature weighting algorithms[1]. We refer to some algorithms using these criteria.

- *Interinstance distance*: this criterion is used in the Kira and Rendell'RELIEF (Kira and Rendell 1992). This method selects a random training case ω_j, a similar positive case ω_a, and a similar negative case ω_b. It then updates the w_i feature weight using:

  $$w_i = w_i - diff(x_{ji}, x_{ai}) + diff(x_{ji}, x_{bi})$$
  where $diff$ is a given metric

  Based on this principle, Kononenko proposes an extension of RELIEF in (Kononenko 1994).
- *Interclass distance*: the average distance between instances belonging to different classes is a good criterion to measure the relevance of a given feature space. However, the use of this criterion is restricted to problems without mutual class overlaps.
- *Probabilistic distance*: in order to correctly treat class overlaps, a better approach consists in measuring distances between probability density functions. This way to proceed often leads to the construction of homogeneity tests (Rao 1965).
- *Class Projection*: this approach assigns weights using conditional probabilities on features that can be indiscriminately nominal, discrete or continuous (Stanfill and Waltz 1986).
- *Entropy*: one can speak about feature selection in terms of information theory. One can then assign feature weights using the Shannon's mutual information (Wettschereck and Dietterich 1995); see also (Koller and Sahami 1996) where the cross-entropy measure is used.

We propose in the next section a new way to proceed using both the contribution of information theory and the rigor of statistical tests. We assume that classifier ability to correctly label instances depends on the existence in the feature space of wide geometrical structures of points identically labelled. We characterize these structures using information contained in the well-known Minimum Spanning Tree. This information is used to apply the following test of *Relative Certainty Gain*.

The Test of Relative Certainty Gain

Geometrical concepts

Our approach is based on the research of characteristics of the learning sample in a neighborhood graph. More precisely we use the *Minimum Spanning Tree*, which is simple to build, and has interesting geometrical properties. The construction of this neighborhood graph allows to exploit local and global information of the concept to learn. We review before some useful definitions about graphs and information measures.

Definition 1 *A tree is a connected graph without cycles.*

Definition 2 *A subgraph that spans all vertices of a graph is called a spanning subgraph.*

Definition 3 *A subgraph that is a tree and that spans all vertices of the original graph is called a spanning tree.*

[1] Feature selection algorithms are weighting algorithms, where irrelevant or weakly relevant features have a zero weight. For more details about feature weighting see (Wettschereck and Aha 1995)

Definition 4 *Among all the spanning trees of a weighted and connected graph, the one (possibly more) with the least total weight is called a Minimum Spanning Tree (MST).*

So, if we have a given metric over the p-dimensional representation space, we can easily build a MST considering the weight of an edge as the distance between its extremities. When features have real values (it is too often an implicit hypothesis in feature selection algorithms!), standard euclidean metric is sufficient. In order to be able to deal with any type of feature (*nominal, discrete, continuous*), our approach requires using specific metrics. In (Wilson and Martinez 1997), new heterogeneous distance functions are proposed, called the Heterogeneous Euclidean-Overlap Metric (*HEOM*), the Heterogeneous Value Difference Distance (*HVDM*), the Interpolated Value Difference Metric (*IVDM*), and the Windowed Value Difference Metric (*WVDM*). These distance functions properly handle nominal and continuous input attributes and allow the construction of a MST in mixed spaces. They are inspired by the Value Difference Metric (Stanfill and Waltz 1986) which defines the distance between two values x and y of an attribute a as following:

$$\begin{aligned} vdm_a(x,y) &= \sum_{c=1}^{k} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^q \\ &= \sum_{c=1}^{k} |P_{a,x,c} - P_{a,y,c}|^q \end{aligned}$$

where

- $N_{a,x}$ is the number of instances in the training set S that have value x for attribute a,
- $N_{a,x,c}$ is the number of instances in S that have value x for attribute a and output class c,
- k is the number of output classes,
- q is a constant, usually 1 or 2,
- $P_{a,x,c}$ is the conditional probability that the output class is c given that attribute a has the value x, i.e., $P(c/x_a)$.

In the case of continuous attributes, HEOM, HVDM, IVDM and WVDM apply different strategies :

- $\frac{|x_a - y_a}{\max_a \quad \min_a}$ for HEOM, where $\max_a$ and $\min_a$ are the maximum and minimum values for attribute a,
- $\frac{|x_a - y_a}{4\sigma_a}$ for HVDM, where σ_a is the standard deviation of the numeric values of attributes a,
- discretization approaches for IVDM and WVDM.

Entropy Notions

Definition 5 *Given*

$$S_k = \left\{ \begin{array}{c} (\gamma_1 .., \gamma_j, .., \gamma_k) \in I\!R^k; \\ \forall j = 1, .., k \ \gamma_j \geq 0 \text{ and } \sum_{j=1}^{k} \gamma_j = 1 \end{array} \right\}$$

the k-dimensional simplex, where k is a positive integer.

An entropy measure is an application from S_k in IR_+, with the following properties (for more details see (Breiman, Friedman and Olshen 1984)): *Symmetry, Minimality, Maximality, Continuity and Concavity*

Definition 6 *The Quadratic Entropy is a function QE from* $[0,1]^k$ *in* $[0,1]$,

$$QE : S_k \rightarrow [0,1]$$
$$(\gamma_1, .., \gamma_k) \rightarrow QE((\gamma_1, .., \gamma_k)) = \sum_{j=1}^{k} \gamma_j(1 - \gamma_j)$$

where k is the number of classes ($k = card(Y)$ with our notations).

Local and Total Uncertainties in the MST

Given the previous definitions, we use the quadratic entropy concept to measure local and total uncertainties in the MST built on the learning set.

Definition 7 *we define the neighborhood* $V(\omega_i)$ *of a given* ω_i *instance belonging to* S *as following:*

$V(\omega_i) = \{\omega_j \in S$ / ω_i *is linked by an edge to* ω_j *in the* $MST\} \cup \{\omega_i\}$

Definition 8 *the local uncertainty* $U_{loc}(\omega_i)$ *for a given* ω_i *instance belonging to* S *is defined as following:*

$$U_{loc}(\omega_i) = \sum_{j=1}^{k} \frac{n_{ij}}{n_{i.}}(1 - \frac{n_{ij}}{n_i})$$
$$\textit{where } n_{i.} = card\{V(\omega_i)\}$$
$$\textit{and } n_{ij} = card\{\omega_l \in V(\omega_i) \mid Y(\omega_l) = y_j\}$$

Definition 9 *the total uncertainty* U_{tot} *in the learning sample is defined as following:*

$$U_{tot} = \sum_{i=1}^{n} \frac{n_{i.}}{n_{..}} \sum_{j=1}^{k} \frac{n_{ij}}{n_{i.}}(1 - \frac{n_{ij}}{n_i})$$
$$\text{where } n_{..} = \sum_{i=1}^{n} n_{i.} = n + 2(n-1) = 3n - 2$$

There are always $n-1$ edges in a MST and each edge is count 2 times (for each of the two extremities).

The Statistical Test

The previous criterion U_{tot} allows to estimate the information level of the learning sample in a given feature space. We propose in this section to provide more than a simple criterion, building a statistical test. In order to correctly estimate feature relevance, a performing approach consists in measuring the class overlap degree of the probability density functions, and compare this one with the degree obtained with a total overlap. This way to proceed consists in applying an homogeneity test, with the following null hypothesis H_0:

$$H_0 : F_1(x) = F_2(x) = ... = F_k(x) = F(x)$$

where $F_i(x)$ is the repartition function of the class i

To be able to apply this test, we must know the law of the statistic used in the test (here, the U_{tot} total uncertainty) under the null hypothesis. Works proposed in (Light and Margolin 1971) show that the distribution of the relative quadratic entropy gain is a χ^2 with $(n-1)(k-1)$ degrees of freedom. Rather than taking directly U_{tot} as variable, we then use the following Relative Certainty Gain,

$$RCG = \frac{U_0 - U_{tot}}{U_0}$$

where U_0 is the uncertainty of the learning set before the construction of the MST.

$$U_0 = \sum_{j=1}^{k} \frac{n_j}{n}\left(1 - \frac{n_j}{n}\right)$$

where $n_j = card\{\omega_i \;/\; Y(\omega_i) = y_j\}$

According to Light and Margolin,

$$n_{..}RCG \equiv \chi^2_{(n-1)(k-1)}$$
$$E(n_{..}RCG) = (n-1)(k-1)$$
$$V(n_{..}RCG) = 2(n-1)(k-1)$$

For reasonably large learning sets ($n > 30$), the distribution of $n_{..}RCG$ is approximately normal with mean $(n-1)(k-1)$ and variance $2(n-1)(k-1)$.

$$n_{..}RCG \approx N((n-1)(k-1), 2(n-1)(k-1))$$

The null hypothesis will then be rejected (with an α risk) if and only if:

$$\frac{n_{..}RCG - (n-1)(k-1)}{\sqrt{2(n-1)(k-1)}} \succ U_\alpha \Longleftrightarrow$$
$$n_{..}RCG > (n-1)(k-1) + U_\alpha\sqrt{2(n-1)(k-1)}$$

where U_α is the value of the repartition function of the normal law $N(O,1)$ at the α risk.

Instead of fixing the α risk in advance (generally 5%), we can calculate the α_c critical threshold necessary for rejecting H_0. Then, we can optimize α_c as an estimation criterion to search for the feature subset which allows to be the farthest from the H_0 hypothesis. Actually, the smaller this risk is, the further from the H_0 hypothesis we are. We use then this α_c risk in the following feature selection algorithm.

The Feature Selection Algorithm

Given p features, $X_1, X_2, ..., X_p$. Among these p attributes, we search for the most discriminant ones using the following algorithm. The heuristic search remains a forward selection, optimizing the critical threshold of the test.

```
1. $\alpha_0 \leftarrow 1$; $E = \emptyset$; $X = \{X_1, X_2, ..., X_p\}$
2. For each $X_i \in X$ do
   Compute the $\alpha_{ci}$ critical threshold in the $E \cup X_i$ feature space
3. Select $X_{\min}$ with $\alpha_{\min} = Min\{\alpha_{ci}\}$
4. If $\alpha_{\min} \ll \alpha_0$ then
   $X = X - \{X_{\min}\}$
   $E = E \cup \{X_{\min}\}$
   $\alpha_0 \leftarrow \alpha_{\min}$
   Return to step 2
   else Return $E$.
```

Set	# feat1	Acc1	# feat2	Acc2
Synt1	10	95.4±2.2	2	96.0±1.3
Synt2	10	58.2±3.6	1	58.2±2.2
Synt3	10	73.2±6.8	5	78.8±1.9
Iris	4	88.5±4.5	2	94.7±3.0
Breast	13	66.2±6.1	3	82.7±4.9
Vote	16	91.5±3.7	1	95.7±2.3
Glass2	9	72.0±5.6	4	73.2±6.3
Xd6	10	78.1±2.9	9	79.9±3.1
Hepatitis	19	82.4±6.0	7	81.9±4.7
EchoCardio	6	72.6±8.0	3	74.8±6.9
Audiology	69	80,1±4.2	21	80,3±3.0

Table 1: Results on synthetic and natural domains: Acc1 corresponds to the accuracy estimates with all the original features (# feat1 is the space size) and Acc2 presents results with the selected feature subset (# feat2 is the subspace size).

Experimental Results

In order to show the interest of a new approach, an experimental study should satisfy two criteria: *relevance* and *insight* (Langley 1998). That's why, in this section, we present some experimental results on two types of problems. The first one concerns synthetic domains. In that case, we know the *a priori* number of relevant and irrelevant features. This way to proceed allows to verify the effects of the algorithm. The second type of problems concerns natural domains. We test our algorithm on 11 datasets, among which 8 belong to the UCI database repository[2]. We run also our algorithm on the three following synthetic problems:

- **Synt 1**: 10 features, among whom 7 are more or less relevant ($V_1 - V_2$ the more relevant), and 3 are irrelevant ones ($V_8 \rightarrow V_{10}$).
- **Synt 2**: 10 features, among whom three are exactly redundant features ($V_1 \rightarrow V_3$), and 7 are irrelevant ($V_4 \rightarrow V_{10}$).
- **Synt 3**: 10 features, including seven identically distributed relevant features ($V_1 \rightarrow V_7$), and 3 irrelevant ($V_8 \rightarrow V_{10}$).

n.b.: For irrelevant features, simulated classes are identically distributed by a normal law $N(0,1)$.

Results of table 1 show that performances of our feature selection algorithm are interesting, eliminating both irrelevant and redundant features. In the majority of cases, the accuracy estimates obtained with a

[2] http://www.ics.uci.edu/~mlearn/MLRepository.html

5-fold-Cross-Validation using a 1-nearest neighbor classifier are better in the selected feature subspace than with the all set, and moreover with smaller standard deviations.

Conclusion

With the development of new data acquisition techniques, and with databases huger and huger, comprehensibility is became very important in machine learning. It is then the duty of algorithms to achieve a high level of performance and explicativity. Feature selection is became a central problem in machine learning. We have presented in this paper a feature selection model based both on information theory and statistical tests. A feature is selected if and only if the information given by this attribute allows to statistically reduce class overlap. Results on synthetic and natural domains show that our statistical tool is suited to treat irrelevant and redundant features, even in very large feature spaces. This is a filter approach which avoids the choice of a given learning method. In future work, we should (i) compare our results with other filter models used for feature selection or feature weighting, and (ii) try to reduce the computational costs linked to the construction of the MST.

References

Aha, D. 1992. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *International Journal of Man-Machine Studies* 36(1):267–287.

Blum, A., and Langley, P. 1997. Selection of relevant features and examples in machine learning. *Issue of Artificial Intelligence.*

Breiman, L.; Friedman, J.; Olshen, R.; and Stone, C. 1984. *Classification And Regression Trees.* Chapman & Hall.

Efron, B., and Tibshirani, R. 1993. *An introduction to the bootstrap.* Chapman & Hall.

John, G.; Kohavi, R.; and Pfleger, K. 1994. Irrelevant features and the subset selection problem. In *Eleventh International Conference on Machine Learning*, 121-129.

Kira, K., and Rendell, L. 1992. A practical approach to feature selection. In *Ninth International Conference on Machine Learning (Aberdeen-Scotland)*, 249–256.

Kohavi, R. 1994. Feature subset selection as search with probabilistic estimates. *AAAI Fall Symposium on Relevance.*

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence* 1137–1143.

Koller, D., and Sahami, R. 1996. Toward optimal feature selection. In *Thirteenth International Conference on Machine Learning (Bari-Italy).*

Kononenko, I. 1995. Estimating attributes: Analysis and extensions of relief. In *1994 European Conference on Machine Learning (Catania Italy)*, 171-182.

Langley, P., and Iba, W. 1993. Average-case analysis of a nearest neighbor algorithm. In *Thirteenth International Joint Conference on Artificial Intelligence*, 889–894.

Langley, P. 1994. Selection of relevant features in machine learning. In *AAAI Fall Symposium on Relavance.*

Langley, P. 1998. Relevance and insight in experimental studies. In *IEEE Expert.*

Light, R., and Margolin, B. 1971. An analysis of variance for categorical date. *Journal of the American Statistical Association (66)* 534–544.

Quinlan, J. 1993. Programs for machine learning. *Morgan Kaufmann.*

Rao, C. 1965. *Linear statistical inference and its applications.* Wiley New York.

Stanfill, C., and Waltz, D. 1986. Toward memory-based reasoning. In *Communications of the ACM*, 1213–1228.

Thrun etal. 1991. The monk's problem: a performance comparison of different learning algorithms. *Technical report CMU-CS 91-197-Carnegie Mellon University.*

Wettschereck, D., and Aha, D. 1995. Weighting features. In *First International Conference on Cased-Based Reasoning ICCBR'95.*

Wettschereck, D., and Dietterich, T. 1995. An experimental comparison of the nearest neighbor and nearest hyperrectangle algorithms. In *Machine Learning, 19*, 5–28.

Wilson, D., and Martinez, T. 1997. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research (6)* 1–34.