Feature Selection for Classification

J. S. Sánchez

Dept. Llenguatges i Sistemes Informàtics
Universitat Jaume I
E-12071 Castelló (Spain)

What Is a Feature?

TRS_DT	TRS_TYP_CD	REF_DT	REF_NUM	CO_CD	GDS_CD	QTY	UT_CD	UT_PRIC
21/05/1993	00001	04/05/1993	25119	10002J	001M	10	CTN	22,000
21/05/1993	00001	05/05/1993	25124	10002J	032J	200	DOZ	1,370
21/05/1993	00001	05/05/1993	25124	10002J	033Q	500	DOZ	1,000
21/05/1993	00001	13/05/1993	25217	10002J	024K	5	CTN	21,000
21/05/1993	00001	13/05/1993	25216	10026H	006C	20	CTN	69,000
21/05/1993	00001	13/05/1993	25216	10026H	008Q	10	CTN	114,000
21/05/1993	00001	14/05/1993	25232	10026H	006C	10	CTN	69,000
21/05/1993	00001	14/05/1993	25235	10027E	003A	5	CTN	24,000
21/05/1993	00001	14/05/1993	25235	10027E	001M	5	CTN	24,000
21/05/1993	00001	22/04/1993	24974	10035E	009F	50	CTN	118,000
21/05/1993	00001	27/04/1993	25033	10035E	015A	375	GRS	72,000
21/05/1993	00001	20/05/1993	25313	10041Q	010F	10	CTN	26,000
21/05/1993	00001	12/05/1993	25197	10054R	002E	25	CTN	24,000

Information Reduction

Prototype Selection

TRS_DT	TRS_TYP_CD	REF_DT	REF_NUM	CO_CD	GDS_CD	QTY	UT_CD	UT_PRIC
21/05/1993	00001	04/05/1993	25119	10002J	001M	10	CTN	22,000
21/05/1993	00001	05/05/1993	25124	10002J	032J	200	DOZ	1,370
21/05/1993	00001	05/05/1993	25124	10002J	033Q	500	DOZ	1,000
21/05/1993	00001	13/05/1993	25217	10002J	024K	5	CTN	21,000
21/05/1993	00001	13/05/1993	25216	10026H	006C	20	CTN	69,000
21/05/1993	00001	13/05/1993	25216	10026H	008Q	10	CTN	114,000
21/05/1993	00001	14/05/1993	25232	10026H	006C	10	CTN	69,000
21/05/1993	00001	14/05/1993	25235	10027E	003A	5	CTN	24,000
21/05/1993	00001	14/05/1993	25235	10027E	001M	5	CTN	24,000
21/05/1993	00001	22/04/1993	24974	10035E	009F	50	CTN	118,000
21/05/1993	00001	27/04/1993	25033	10035E	015A	375	GRS	72,000
21/05/1993	00001	20/05/1993	25313	10041Q	010F	10	CTN	26,000
21/05/1993	00001	12/05/1993	25197	10054R	002E	25	CTN	24,000

Feature Selection

3

Different Views

- Selection: a process to obtain a subset from the initial feature set.
- Weighting: a process to assign a weight to each original feature.
- Extraction: a process to create a new feature set by transforming or combining the original attributes.

4

Definition of Feature Selection

- It involves picking up a subset of "relevant" (and "non-redundant") features.
- As a by-product, it also obtains a reduction of the feature space dimensionality.

Relevance of Features

- Different degrees of relevance in terms of Bayes rule:
  - Strongly relevant features.
  - Weakly relevant features.
  - Irrelevant features.

Strong Relevance

- A feature *A* is strongly relevant if removal of *A* alone will result in performance deterioration of an optimal Bayes rule.

7

Weak Relevance

- A feature *A* is weakly relevant if it is not strongly relevant, but in some contexts may contribute to prediction accuracy of an optimal Bayes rule.

8

Irrelevance

- A feature *A* is irrelevant if it is not either weakly relevant or strongly relevant.

9

Goals of Feature Selection

- Attempts to pick up the minimally sized subset of features according to several criteria:
  - Classification accuracy should not decrease significantly.
  - Resulting class distribution derived from the selected subset should be as close as possible to the original class distribution (separation between classes).

10

Mathematical Formulation

- Feature selection is a combinatorial optimization problem:
  - *Given a set Y of D different features, select a subset $X \subseteq Y$ of size d that optimizes a certain objective function J(X).*

$$J(X) = \max_{\substack{Z \subseteq Y \\ |Z| = d}} J(Z)$$

11

Trivial Solution

- To perform an exhaustive search for the best subset with $d \leq D$ features.
  - To test all possible subsets of size *d*.

$\binom{D}{d}$ combinations !!!

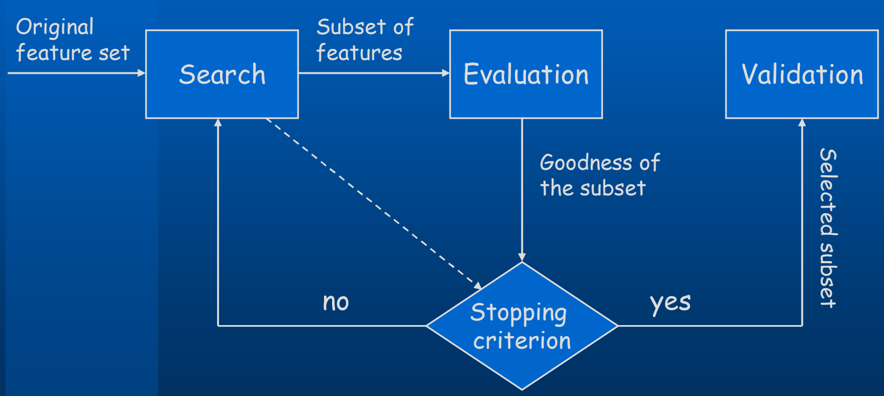
12

An Alternative

- Feature selection algorithms:
  - Based on using a search strategy to define a candidate subset of attributes and a certain objective function to evaluate the goodness of the selected subset.

13

Feature Selection Process (I)



14

Feature Selection Process (II)

- Search procedure: to generate next candidate feature subset.
- Evaluation function: to asses the quality of the subset under examination.
- Stopping criterion: to decide when to stop.
- Validation: to check whether the subset is valid.

Search Procedures

- Optimal solution:
  - Exhaustive (complete) search.
  - Branch & bound search.
- Suboptimal solution:
  - Sequential search.
  - Random search.

Exhaustive Search

- Examines all combinations of feature subset.
- Order of the search space $O(2^D)$.
- Optimal subset is achievable.
- Too expensive if feature space is large.

17

Sequential Search

- Selection is directed under certain guideline
- Incremental generation of subsets.
- Search space is smaller and faster in producing results.

18

Sequential Search Algorithms

- Different algorithms depend on how to start and how the subsequent operations perform:
  - Forward Sequential Search (FSS).
  - Backward Sequential Search (BSS).
  - Sequential Floating Search (SFS).

19

FSS Algorithm

- It starts with an empty feature set.
- At each iteration, the "best" feature is added to the set.

20

BSS Algorithm

- It starts with a set containing all available features.
- At each iteration, it removes the feature whose removal yields the maximal performance improvement.

21

Drawbacks of FSS and BSS

- They cannot correct previous inclusions or removals .
- It is possible to obtain non-optimal solutions.

22

SFS Algorithms

- Two variants: forward (FSFS) and backward (BSFS).
- An improvement over FSS and BSS by means of the "conditional" inclusion (or deletion) of attributes.

SFS Algorithms (cont.)

- After each iteration, we check whether the current feature combination is the "best" subset. If not, the attribute included (or excluded) is now removed (or added).

Random Search

- No predefined way to select feature candidate.
- Picks features at random.
- Optimal subset depends on the number of trials.
- Requires more user-defined input parameters: optimality will depend on how these parameters are defined.

25

Evaluation Function

- Optimal subset is always relative to a certain evaluation (or objective) function.

26

Different Evaluation Functions

Used by a filter method

- Distance measure.
- Information measure.
- Dependency measure.
- Consistency measure.

Used by a wrapper method

- Classifier error rate.

27

Feature Selection Methods

- Filter approaches: they "filter" (select) the irrelevant attributes before learning occurs.
- Wrapper approaches: they use the final learning algorithm as their evaluation (objective) function.

28

Filter Approaches

Set of candidate input variables → Subset selection algorithm → Learning algorithm

- It ignores the effect of the selected subset on the performance of the classifier.

29

Distance Measure

- Euclidean distance: $z^2 = x^2 + y^2$
- Concept: select those features that support instances from the same class to stay within the same proximity.
- Instances from the same class should be closer in terms of distance than those from different classes.

30

Information Measure

- Determines the information gain due to each attribute.
- Information gain can be measured by means of entropy.

31

Entropy

- Determines the "impurity" of an instance set.
- It is maximum when all classes are represented by the same proportion of instances.

32

Joint Entropy

$$E(S) = -\sum_{i=1}^{c} p_i \log_2(p_i)$$

where p_i is the percentage of instances from class i in the subset S.

$$p_i = \frac{|S_i|}{|S|}$$

33

Information Gain

- From entropy, the information gain for a feature A can be defined as:

$$G(s, A) = E(S) - \sum_{i=1}^{d} \frac{|S_i|}{|S|} E(S_i)$$

- Then, feature A is selected over another B if $G(S,A) > G(S,B)$.

34

Dependency Measure

- Determines the correlation between a feature and a class label.
- Correlation coefficient (in the range 0-1) measures the degree of linear relation between two variables. A value equal to 0 means that there is no relation between them.

35

Degree of Redundancy

- To detect redundant features, we can measure the correlation coefficient between attributes *X* and *Y*.

$$r^2 = \frac{s_{XY}^2}{ss_{XX}\, ss_{YY}}$$

$$ss_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad ss_{YY} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad ss_{XY} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

36

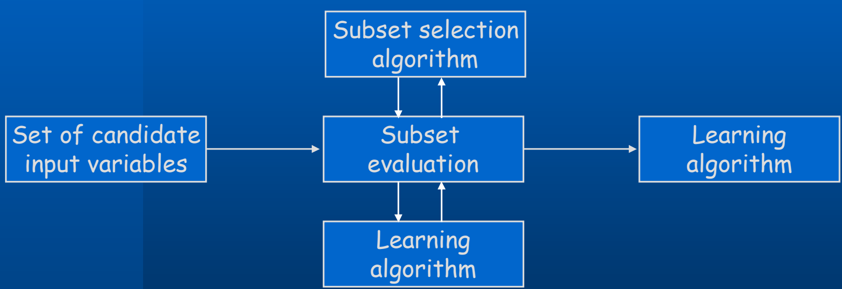
Consistency Measure

- Finds out the minimal subset which satisfies the acceptable inconsistency rate that is usually set by the user.

	f_1	f_2	*class*
First instance	0	1	A
Second instance	0	1	B

inconsistent

Wrapper Approaches



Classifier Error Rate

- Used as an evaluation measure in the wrapper approach.
- If (error rate with feature subset X < predefined threshold value) then select feature subset X.
- High accuracy but computationally very expensive.
- Loss of generality.

39

Some Algorithms

Measures	Search		
	Heuristic	Complete	Random
Distance	Relief	Branch & Bound	
Information	Decision Tree Method	Minimal Description Length Method	
Dependency	Probability of Error & Average Correlation Coefficient Method		
Consistency		Focus	Las Vegas
Classifier Error Rate	BSS, FSS, FSFS, BSFS	AMB & B	LVW

40

Definition of Feature Weighting

- It weights, instead of selecting, each attribute in the original feature set.
- It aims at decreasing the classifier error rate, not at reducing the computational cost

41

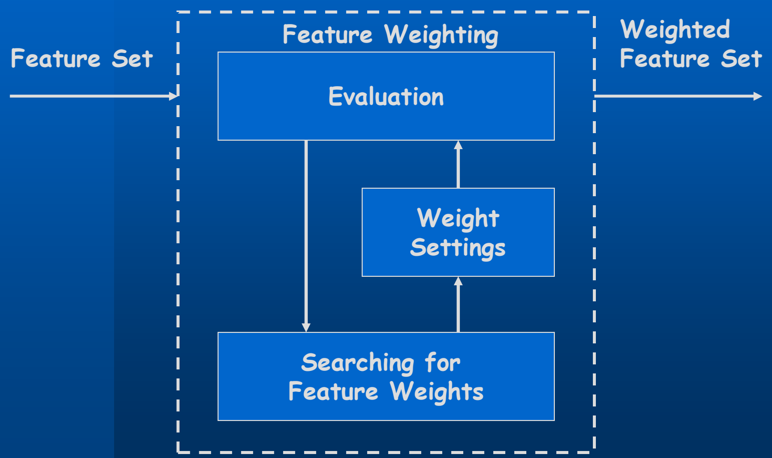
The Weighting Problem

- To map the original set Y with D attributs into a new set X of D features with different weight:

$$X = \{w_1y_1, w_2y_2, \ldots, w_Dy_D\}$$

42

Feature Weighting Process



43

Rationale of Feature Weighting

- Irrelevant features represent a little influence on the classification of new patterns.
- Then, the solution may consist of assigning weights to attributes according to their contribution to the problem.

44

Weighting vs. Selection

- Feature selection constitutes a restricted case of feature weighting:
  - Feature selection algorithms assign binary weights to features: 0 (minimum relevance) and 1 (maximum relevance).

45

Weighting Strategies

- The weighting strategies try to:
  - reward those attributes providing correct classifications.
  - punish those attributes providing wrong classifications.

46

Some Weighting Strategies

- To set feature weights according to the result of predictions.
- To set feature weights according to the class of the nearest neighbours.
- To set feature weights according to the class conditional probabilities.

47

Conclusions

- Different problems related to attributs: irrelevant features, redundant features, harmful features.
- Different perspectives: selection, weighting, extraction.

48

Conclusions (cont.)

- No single method works well under all conditions:
  - Some can handle noise, but not redundant or correlated features.
  - Some can detect redundant features, but not when data is noisy.

49

Conclusions (cont.)

- Finding a good feature subset is an important problem for real datasets. Thus, a good subset can:
  - Simplify data description.
  - Reduce the task of data collection.
  - Improve accuracy and performance.

50