Feature Selection for Classification

Data preprocessing and Feature Selection

MOSAIC PINV15-257

Miguel García Torres
mgarciat@upo.es
Lenguajes y Sistemas Informáticos, Universidad Pablo de Olavide
Intelligent Data Analysis research group

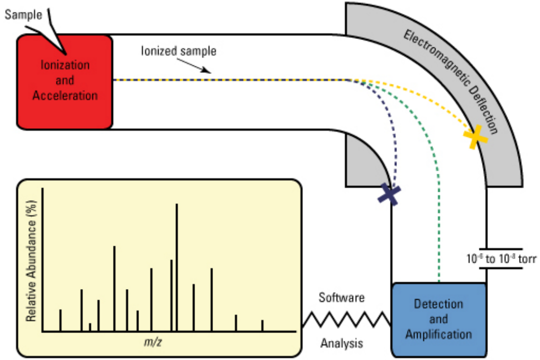
CONACYT

Outline

CONACYT

Proteomic mass spectrometry analysis

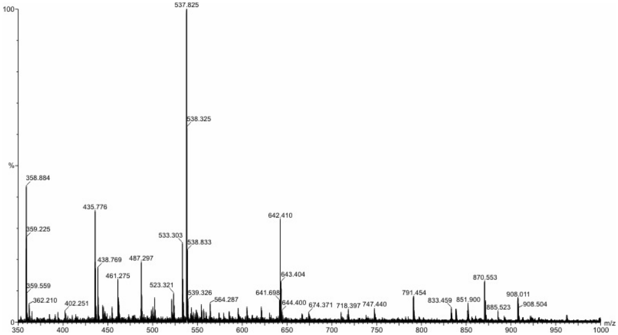
- Enhanced data acquisition ⇒ large high-dimensional data!
- MS analysis ⇒ Peptide and protein analysis.
- Objective: Biomarker detection.



Proteomic mass spectrometry analysis

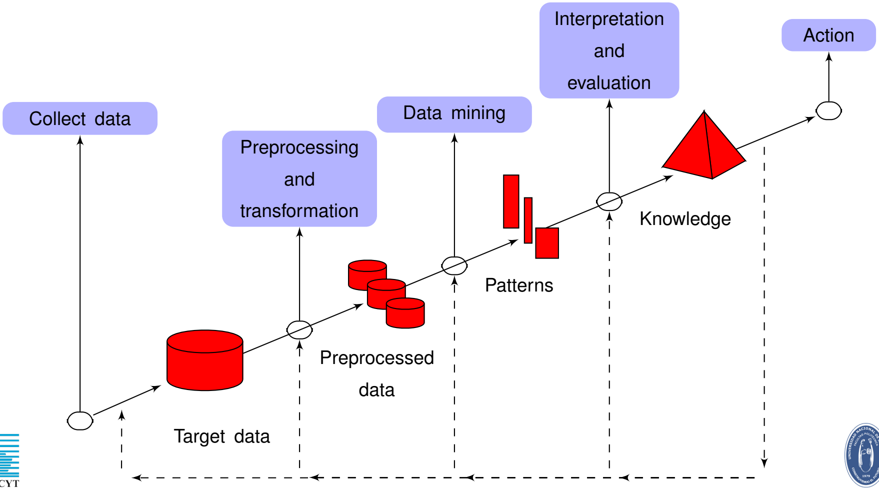
Challenges

- Small high-dimensional dataset.
- Orignal signal decomposition unknown.
- No standard data preprocessing workflow.

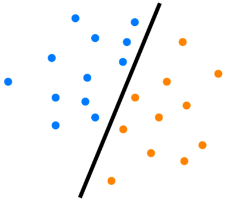


CONACYT

The Knowledge Discovery in Databases process



Classification



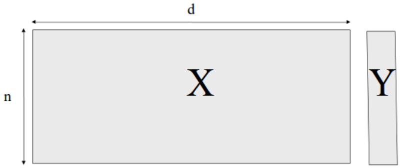
Notation

- $X = \{X_j, j = 1, \ldots, d\}$ full set of features.
- $Y \equiv$ the class variable (target class to be learned).
- $E = (\mathbf{x}, y) \equiv$ training set.
- $T = (\mathbf{x}, ?) \equiv$ test set.

Given E, the aim of classification is to learn a function

$$\mathcal{C} : X \to Y.$$

Feature selection for classification



- Not all the features are equally useful $\Rightarrow$ removing some of them may improve the predictive model $\mathcal{C}$.

$\Rightarrow$ The objective of feature selection is to find the subset of features $S \in X$ with which $\mathcal{C}$ achieves the lowest error rate.

Benefits

- Reduction in the cost of acquisition of the data.
- Improvement of the comprehensibility of the model.
- Faster induction of the final classification model.
- Improvement in classification accuracy.

Feature selection (FS)

- FS traditionally focused on finding a highly discriminating power set of features for minimizing the classification error rate.
- Several works have made an effort for defining the different feature types according to their contribution to the meaning of the class concept.
- In this context, feature relevance has arisen as a measure of the amount of relevant information that a feature may contain about the class in classification tasks.
- A feature is considered irrelevant if it contains no information about the class and therefore it is not necessary at all for the predictive task
- Relevant features are those that embody information about the class concept

CONACYT

Feature selection (FS)

- Optimal feature subset defined with respect to the induction algorithm:
  Given an inducer $\mathcal{I}$*, and a training dataset* E *with features* $X_1, \ldots, X_d$*, from a distribution* $\mathcal{D}$ *over the labeled instance space, an* **optimal feature subset,** S_{opt}**,** *is a subset of the features such that the accuracy of the induced classifier* $\mathcal{C} : \mathcal{I}(\mathcal{D})$ *is maximal.*
- Optimal feature subset not necessarily unique.
- Problem: distribution of data unknown.

$\Rightarrow$ Accuracy of the classifier must be estimated from data.

CONACYT

Kohavi & John [19]

$S_j = X \setminus \{X_j\}$.

- **Strong relevance** $\equiv$ A feature X_j is strongly relevant iff

$$P(Y|X_j, S_j) \neq P(Y|S_j).$$

- **Weak relevance** $\equiv$ A feature X_j is weakly relevant iff

$$P(Y|X_j, S_j) = P(Y|S_j).$$

  and $\exists\ S_j'$, such that

$$P(Y|X_j, S_j') \neq P(Y|S_j).$$

- **Irrelevance** $\equiv$ A feature X_j is irrelevant iff

$$\forall\ S_j' \subseteq S_j, P(Y|X_j, S_j') = P(Y|S_j).$$

Target concept

$$Y = X_1 \oplus X_2.$$

where
$X_4 = \bar{X}_2$
$X_5 = \bar{X}_3$

X_1	X_2	X_3	X_4	X_5	Y
0	1	1	0	0	0
0	1	0	0	1	0
0	0	1	1	0	1
0	0	0	1	1	1
1	1	1	0	0	1
1	1	0	0	1	1
1	0	1	1	0	0
1	0	0	1	1	0

- $X_1 \equiv$ strongly relevant.
- $X_2, X_4 \equiv$ weakly relevants.
- $X_3, X_5 \equiv$ irrelevants.
- model with highest accuracy $\{X_1, X_2\}, \{X_1, X_4\}$

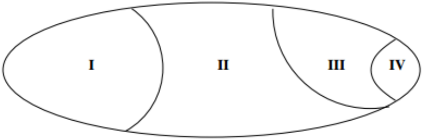
CONACYT

Feature redundancy

- Feature redundancy is usually presented in terms of feature correlation.
- Perfectly correlated features are truly redundant in the sense that no additional information is gained by adding them.
- Redundancy may exist between two uncorrelated features.
- Two highly correlated features may improve the accuracy $\Rightarrow$ correlation cannot be adequated to feature redundancy.

CONACYT

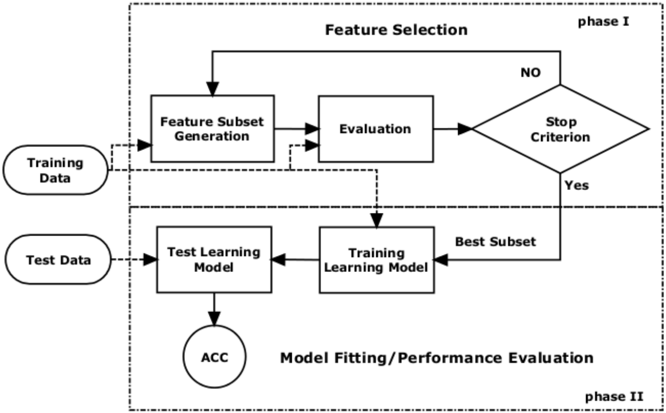
- Selecting the most relevant variables is usually suboptimal for building a predictor, particularly if the variables are redundant
- A subset of useful variables may exclude many redundant, but relevant, variables.
- Relevance does not imply optimality $\equiv$ Let X_1, X_2, X_3 be binary features. Let the distribution of instances be uniform, and assume that the target concept is $J(X_1, X_2, X_3) = (X_1 \wedge X_2) \vee X_3$. In this case, all features are relevant but the optimal subset of features is $\{X_3\}$.



I: Irrelevant features
II: Weakly relevant and redundant features
III: Weakly relevant but non–redundant features
IV: Strongly relevant features
III + IV: Optimal subset

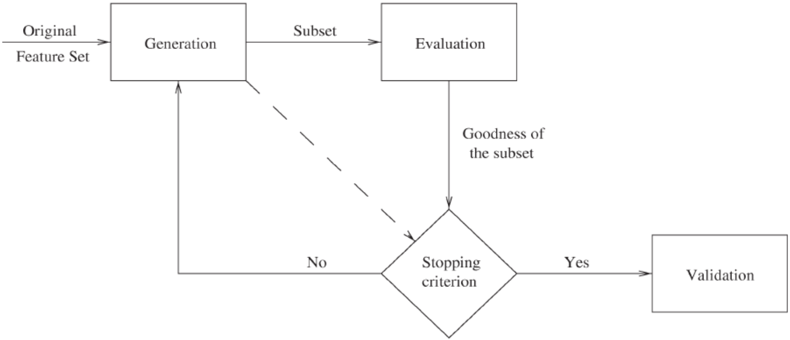
CONACYT

Feature selection workflow



CONACYT

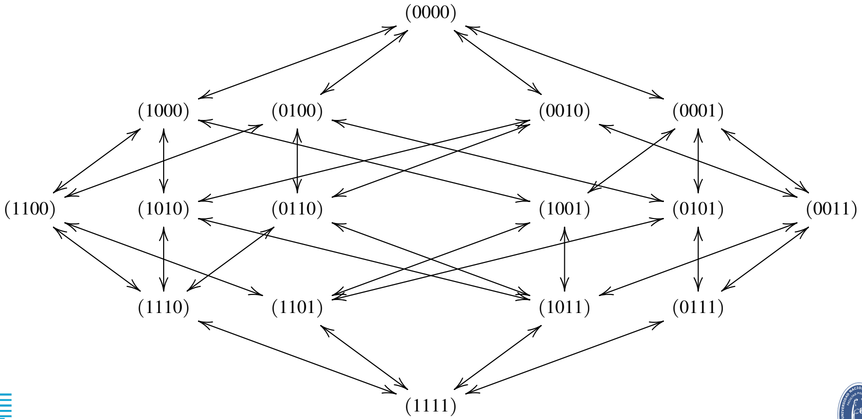
Steps in a typical feature selection method



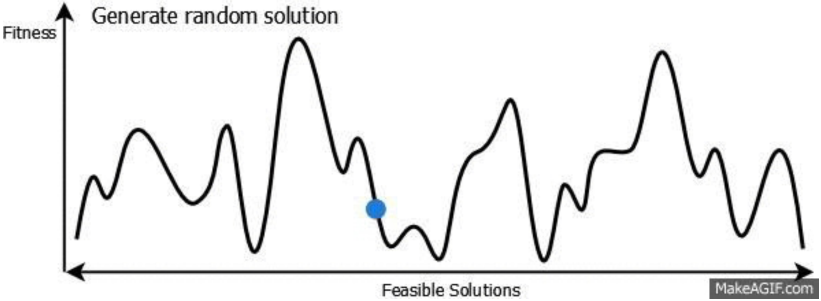
- Feature subset generation $\equiv$ select subset candidate.
- Evaluation $\equiv$ compute relevancy value of the subset.
- Stopping criterior $\equiv$ determine whether subset is relevant.
- Validation $\equiv$ verify subset validity.

Feature subset generation

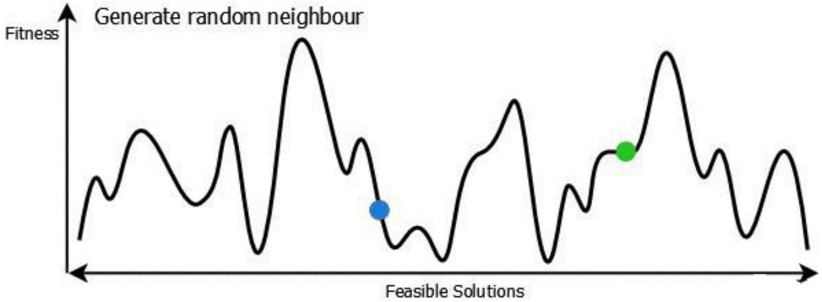
Search space



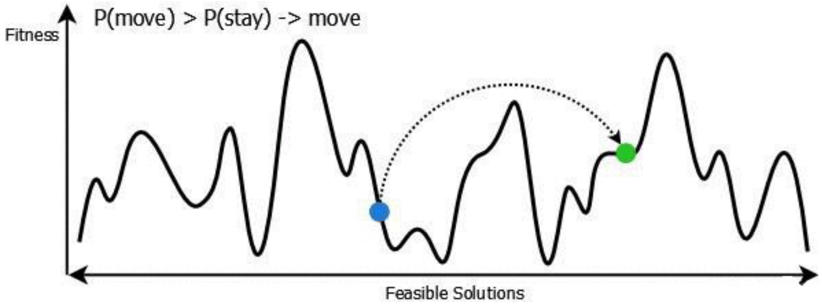
Search space



Search space



Search space



Feature subset generation

Approaches to examine the search space

- Complete $\equiv$ it does a complete search for the optimal subset according to the evaluation function.
  $\Rightarrow$ Worst case: Exhaustive search ($\mathcal{O}(2^d)$).
  $\Rightarrow$ Optimality of the feature subset, according to the evaluation funcion, is guaranteed.
- Heuristic $\equiv$ it generates the subsets under certain guidelines.
  $\Rightarrow$ Optimality is not guaranteed.
  $\Rightarrow$ Procedures very simple to implement and fast in producing results.
  $\Rightarrow$ Search space is usually quadratic ($\mathcal{O}(d^2)$).
  - Deterministic $\equiv$ it generates the subsets in a predefined way.
  - Non deterministic $\equiv$ it generates the subsets randomly.

Evaluation

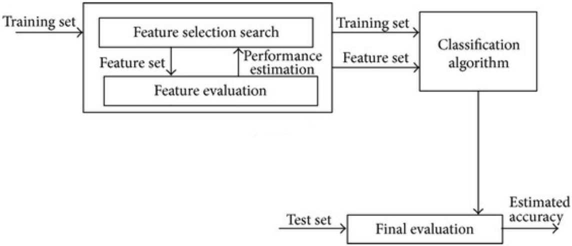
Determines the relevancy of the generated feature subset candidate towards the classification task.

Type of evaluation functions

- Filter
  1. Distance (euclidean distance, Manhattan distance, etc.).
  2. Information (entropy, informacion gain, etc.)
  3. Dependency (correlation).
  4. Consistency (min-features bias).
- Wrapper (classifier).

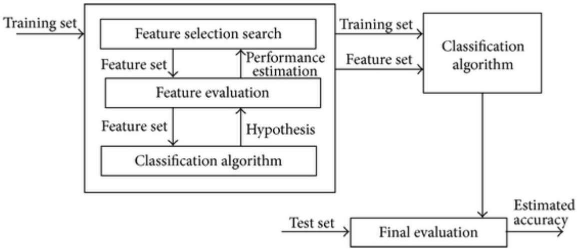
CONACYT

Filter approach



- FS done as a preprocessing step.
- Subsets evaluated according to intrinsic properties of the data.
- Computationally fast $\Rightarrow$ can be scaled to high-dimensional datasets.
- Drawback: Effect of FS on induction algorithm is not known.

Wrapper approach



- Wrappers use the learner as a black box to score the subsets of features according to their predictive power.
- The quality of feature subsets for classification is defined with respect to the induction algorithms.
- Drawback: Wrappers are slow.

Filters vs. wrappers

- Wrappers tend to have higher risk of overfitting than filters.
- Filter may lead to worse accuracy than wrappers.
- Filters are independent of the learner $\Rightarrow$ FS done once for a given training dataset.

Categorization of feature selection methods

Search **Evaluation**	Complete		Heuristic			
			deterministic		non deterministic	
Distance	*B&B*	[31]	*Relief*	[18]		
	BFF	[42]	*ReliefF*	[22]		
	*Seg*84	[36]				
	EUBAFES	[34]				
Information	*MDLM*	[37]	*SFG*	[4]	*PGVNS* [12]	
			DT − CBL	[3]		
			DTM	[4]		
			*KS*96	[20]		
			FCBF	[43]		
			MIFS	[2]		
			CR	[40]		
Dependence			*POE + ACC*	[30]		
			PRESET	[28]		
Consistency	*FOCUS*	[1]	*SetCover*	[6]	*LVF*	[26]
	*Sch*93	[35]	*VCC*	[41]	*SLV*	[27]
	*MIFES*1	[32]			*QBB*	[7]
	ABB	[24]				
Error	*AMB&B*	[11]	**SFS**	[8]	*LVW*	[25]
	BS	[9]	**SBE**	[8]	*GA*	[38]
	LC	[14]	*SBE − SLASH*	[5]	*SA*	[9]
	BC	[15]	*SFFS*	[33]	*FSSEBNA*	[16]
	PQSS	[9]	*BDS*	[9]	**SS** [12]	
			RACE	[29]		
			RC	[10]		
			RACE	[29]		
			Oblivion	[23]		
			IS	[39]		
			RFE	[13]		

CONACYT

Sequential Forward Selection (SFS)

Main idea

- Deterministic heuristic search.
- Filter and wrapper approach.
- Complexity ($\mathcal{O}(d^2)$).
- Forward search.
- Starts with empty set.
- Each step, adds the best feature if its addition improves current solution.
- SFS performs best when the optimal subset is small.
- SFS is unable to remove features $\Rightarrow$ the solution can get stuck in a local optimum.

Pseudocode

Procedure *Sequential Forward Search*

begin

$S \leftarrow \{\emptyset\}$

repeat

  foreach $X_j \notin S$;

    $J_j \leftarrow J(S \cup \{X_j\})$;

  Let $j' \leftarrow arg\, \max\{J_j\}$;

  $S' \leftarrow S \cup \{X_{j'}\}$;

  if $J(S') > J(S)$ **then**

   $S \leftarrow S'$;

   $J(S) \leftarrow J(S')$;

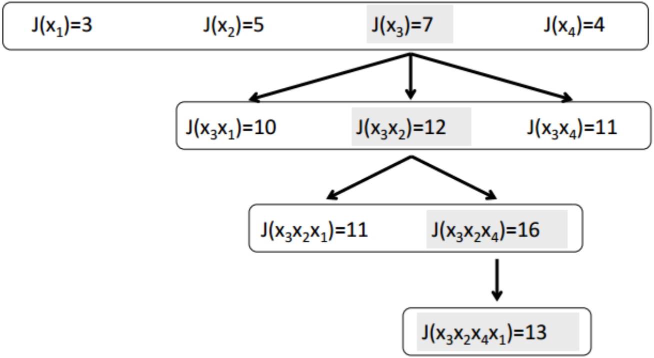
until $(J(S') \leq J(S) \,||\, |S'| == d)$

end

Example od execution of SFS

Objective function

$$J(X) = -2x_1X_2 + 3x_1 + 5x_2 - 2x_1x_2x_3 + 7x_3 + 4x_4 - 2x_1x_2x_3x_4.$$



CONACYT

Sequential Backward Elimination (SBE)

Main idea

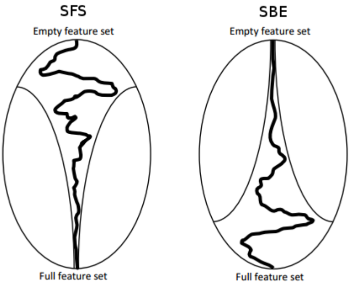
- Deterministic heuristic search.
- Filter and wrapper approach.
- Complexity ($\mathcal{O}(d^2)$).
- Starts with the full set set.
- Each step, removes the worst feature if its removal improves current solution.
- SBE performs best when the optimal subset is large.
- It is unable to reevaluate the usefulness of a feature after it has been discarded.

CONACYT

Pseudocode

Procedure *Sequential Backward Elimination*
begin
$S \leftarrow \{X_1, \ldots, X_d\}$
repeat
  foreach $X_j \in S$;
    $J_j \leftarrow J(S \setminus \{X_j\})$;
  Let $j' \leftarrow arg\, \max\{J_j\}$;
  $S' \leftarrow S \setminus \{X_{j'}\}$;
  if $J(S') > J(S)$ **then**
    $S \leftarrow S'$;
    $J(S) \leftarrow J(S')$;
until $(J(S') \leq J(S) \,||\, |S'| == 1)$
end

SFS vs. SBE



- $|S_{SFS}| \leq |S_{SBE}|$.
- SFS may suffer of overfitting.
- $t_{SFS} \leq t_{SBE}$.
- SBE cannot be applied to medium high-dimensional data.
- Complexity ($\mathcal{O}(d^2)$) $\Rightarrow$ not suitable for large high-dimensional data.

CONACYT

Fast Correlation Based Filter (FCBF)

Main idea

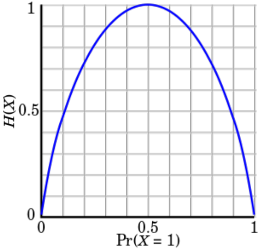
- Deterministic heuristic search.
- Filter approach $\equiv$ information theory measures.
- Complexity:
  - best case: only one feature selected ($\mathcal{O}(d)$).
  - worst case: all features are selected ($\mathcal{O}(d^2)$).
- Two steps:
  1. Analysis of relevance.
  2. Analysis of redundance.

CONACYT

Definitions
Entropy

It measures the uncertainty about the value of a random variable X.

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i)).$$



- Feature X with values $\{0, 1\}$.
- Entropy is 0 if there is no uncertainty.

Definitions
Conditional entropy

It measures the uncertainty about the value of X given the value of Y.

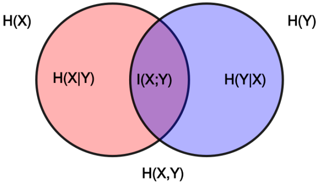
$$H(X|Y) = -\sum_j P(y_j) \sum_i P(x_i|y_j) \log_2(P(x_i|y_j)).$$

- $H(X|Y) = 0$ iff the value of X is completely determined by the value of Y.
- $H(X|Y) = H(X)$ iff X and Y are independent.

Definitions
Information Gain

It measures the reduction in uncertainty about the value of X given the value of Y

$$IG(X;Y) = H(X) - H(X|Y).$$

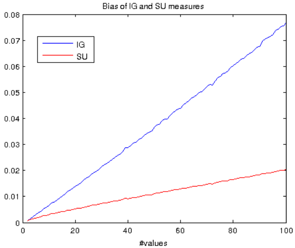


- $H(X) \equiv$ circle on the left (red and violet).
- $H(Y) \equiv$ circle on the right (blue and violet).
- $H(X,Y) \equiv$ area contained by both circles.
- $H(X|Y) \equiv$ red.
- $H(Y|X) \equiv$ blue.
- $I(X;Y) \equiv$ violet.

Definitions

Symmetrical Uncertainty (SU)

$$SU(X,Y) = 2\left[\frac{IG(X;Y)}{H(X)+H(Y)}\right].$$



- 1000 examples generated randomly.
- 99 features with:
  - $X_1 = \{0,1\}$,
  - $X_2 = \{0,1,2\}$,
  - $\ldots$,
  - $X_{99} = \{0,1,\ldots,99\}$.
- Target class generated randomly $\Rightarrow$ MI and SU values should be close to 0,

CONACYT

Definitions

Approximate Markov blanket (AMb)

Given two features X_i and X_j $(i \neq j)$ so that $SU(X_j, \mathcal{Y}) \geq SU(X_i, \mathcal{Y})$, then X_j forms an approximate Markov blanket for X_i iff $SU(X_i, X_j) \geq SU(X_i, \mathcal{Y})$.

Predominant feature

Given a set of features S, a relevant feature is a predominant feature iff it does not have any AMb in S.

CONACYT

Analysis of relevance

- Relevance measure $\equiv$ Symmetrical Uncertainty $SU(X_j, Y), j = 1, \ldots, d$.
- Given δ, a feature X_j is irrelevant if $SU(X_j, Y) \leq \delta$.

Analysis of redundance

Markov blanket [21] $\equiv$ Given a feature X_j, $M_j \subset X$ ($X_j \notin M_j$) is said to be a Markov blanket for X_j iff
$P(X - M_j - \{X_j\}, \mathcal{Y}|X_j, M_j) = P(X - M_j - \{X_j\}, \mathcal{Y}|M_j)$.

- M subsumes not only the information that X_i has about $\mathcal{Y}$ but also about all of the other features.
- A feature $X_j \in S$ is redundant and, so, it can be removed from S if we find a Markov blanket M for X_j within S.

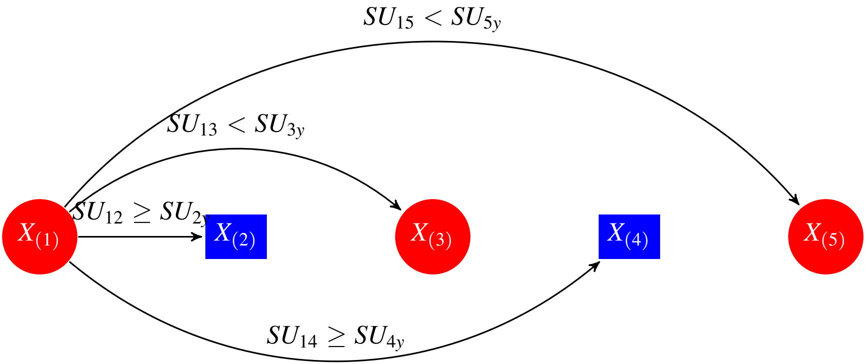
Analysis of redundance

- Features are ordered in descending order according to the SU values.
- First feature $X_{(1)}$ is a *predominant feature*.
- Second iteration $\Rightarrow$ remove those features $X_{(j)}$ for which $X_{(1)}$ is an AMb.
- Second iteration $\Rightarrow$ select next non-removed feature as *predominant feature* and remove thoses features for which, this feature forms an AMb.
- So on.

CONACYT

Example of analysis of redundance

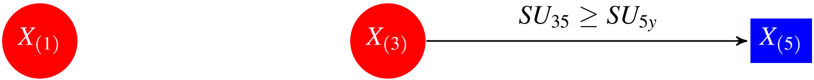
Iteration 1



CONACYT

Example of analysis of redundance

Iteration 2



CONACYT

Pseudocode

Procedure *Fast Correlation Based Filter*
begin
for $i = 1$ to d **do**
  calculate SU_{ic} for X_i;
  if $(SU_{ic} > \delta)$
    append X_i to S'_{list};
end;
order S'_{list} in descending SU_{ic} value;
$X_j =$ getFirstElement(S'_{list});
do begin;
  $X_i =$ getNextElement(S'_{list}, X_j);
  if($X_i <> NULL$)
    do begin;
      if($SU_{ij} \geq SU_{ic}$);
        remove X_i from S'_{list};
      $X_i =$ getNextElement(S'_{list}, X_j);
    end until ($X_i == NULL$);
  $X_j =$ getNextElement(S'_{list}, F_j);
end until ($X_j == NULL$);
$S_{best} = S'_{list}$;
end

Scatter Search (SS)

Main idea

- Heuristic and non deterministic method.
- Population based strategy.
- Evolution based on intensification and diversification strategies.

CONACYT

SS pseudocode

```
Procedure Scatter search
begin
    GeneratePopulation (InitPop);
    GenerateReferenceSet (RefSet);
    repeat
        repeat
            SelectSubset (Subset);
            CombinationMethod (Subset, CurSol);
            ImprovementMethod (CurSol, ImpSol);
        until (StoppingCriterion_1)
        UpdateReferenceSet (RefSet);
    until (StoppingCriterion_2)
end
```

Generate initial population

- Let L be an ordered subset with featuresofel subconjunto formado por los atributos con mayor poder predictivo (tal que $J(\{x_j\}) \geq J(\{x_{j+1}\})$).

Procedure *Generate initital population*
{
$S \leftarrow \emptyset$;
Order $\{X_j\}, j = 1, \ldots, d$ such that $f(X_j) \geq J(X_{j+1})$;
$L \leftarrow \{X_j\}, j = 1, \ldots, k$ such that $k \leq d$;
repeat
  Select randomly $X_{j*} \in L$;
  if $J(\{X_{j*}\} \cup S) \geq J(S)$
    $S \leftarrow S \cup \{X_{j*}\}$
    $L \leftarrow (L \setminus \{X_{j*}\}) \cup \{X_j\}, X_j \notin L$
until $J(\{X_{j*}\} \cup S) < J(S)$
}

Update the Reference Set

- $|RefSet| = |RefSet1| + |RefSet2|$.
- $RefSet1 \equiv$ quality and $RefSet2 \equiv$ diversity.

Procedure *Update the Reference Set*
{
$RefSet \leftarrow \emptyset$
$RefSet \leftarrow |RefSet|$ best solutions from Pop.
Let $C = \cup_{X_j \in RefSet} X_j$
repeat $\forall\ S \notin RefSet$
  Calculate $Div(S, C) = |(S \cup C) \setminus (S \cap C)|$
  Let $S^* \leftarrow \arg\max Div(S, C) : S \notin RefSet$.
  $RefSet \leftarrow RefSet \cup S^*$.
  $|RefSet| \leftarrow |RefSet| + 1$.
  Update C
until $|RefSet| = |RefSet1| + |RefSet2|$
}

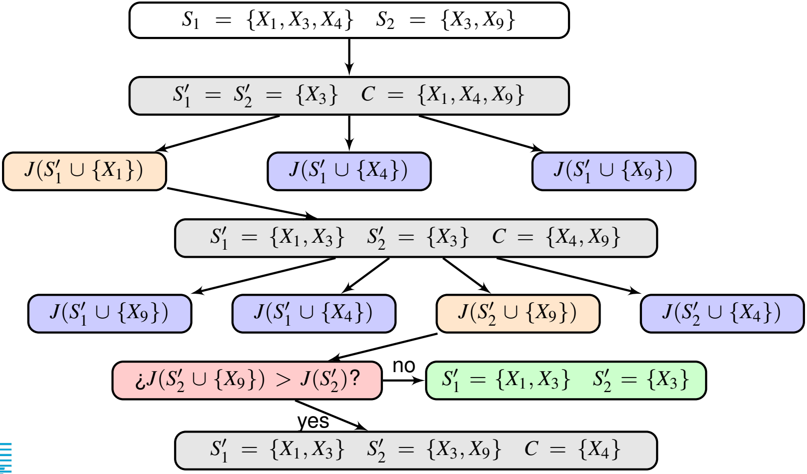
Combination method

- $S_i \equiv$ Solution i.
- $S'_i \equiv$ new solution generated i.
- $(S_1, S_2) \rightarrow (S'_1, S'_2)$.

```
Procedure Greedy Combination
{
S'_1 = S'_2 ← S_1 ∩ S_2, C = (S_1 ∪ S_2) \ (S_1 ∩ S_2).
S'_1 ← S'_1 ∪ {X_{j*}} : X_{j*} = max_j{J(S'_1 ∪ {X_j})}.
repeat
   j*_k : J(S'_k ∪ {X_{j*_k}}) = max_j{J(S'_k ∪ {X_j})}, k = 1, 2;
   Let j** = max_k{J(S'_k ∪ {X_{j*_k}})}
   if J(S'_k ∪ {X_{j**}}) > J(S'_k)
     S'_k ← S'_k ∪ {X_{j**}}
     C ← C \ {X_{j**}}
until J(S'_k ∪ {X_{j**}}) ≤ J(S'_k), k = 1, 2
}
```

CONACYT

Example of the combination method



CONACYT

Improvement method

- Let $CA = \{X_j : X_j \notin S\}$, ordered according to the evaluation method ($J(\{x_j\}) \geq J(\{x_{j+1}\})$).

Procedure *Improvement method*
{
$j \leftarrow 0$
repeat
  $j \leftarrow j + 1;$
  if $J(S \cup \{X_j\}) \geq J(S)$
    $S \leftarrow S \cup \{X_j\}$
until $j \leftarrow |CA|$
}

More topics related to feature selection

- Stability of the FS strategies.
- FS applied to regression and clustering.
- Causal Feature Selection.

[1] H. Almuallim and T. G. Dietterich.
Learning with many irrelevant features.
In A. P. . T. M. Press, editor, *Proceedings of the Ninth National Conference on Artificial Intelligence*, pages 547–552, 1992.

[2] R. Battiti.
Using mutual information for selecting features in supervised neural net learning.
IEEE Transactions on Neural Networks, 5(4):537–550, 1994.

[3] C. Cardie and N. Howe.
Empirical methods in information extraction.
In D. Fischer, editor, *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 65–79. Morgan Kaufmann, 1997.

CONACYT

[4] C. Cardie93.
Using decision trees to improve case-based learning.
In *Proceedings of the Tenth International Conference on Machine Learning*, pages 25–32, 1993.

[5] R. Caruana and D. Freitag.
Greedy attribute selection.
In *Proceedings of International Conference on Machine Learning*, pages 28–36. AAAI Press / The MIT Press, 1994.

[6] M. Dash.
Feature selection via set cover.
In *Proceedings of IEEE Knowledge and Data Engineering Exchange Workshop*, pages 165–171, 1997.

[7] M. Dash and H. Liu.
Hybrid search of feature subsets.
In *Pacific Rim International Conference on Artificial Intelligence*, pages 238–249, 1998.

CONACYT

[8] P. A. Devijver and J. Kittler.
Pattern Recognition: A Statistical Approach.
Prentice Hall International, 1982.

[9] J. Doak.
An evaluation of feature selection methods and their application to computer security.
Technical report, University of California, Department of Computer Science, 1992.

[10] P. Domingos.
Contextsensitive feature selection for lazy learners.
Artificial Intelligence Review, 11:227–253, 1997.

[11] I. Foroutan and J. Sklansky.
Feature selection for automatic classification of non-gaussian data.
IEEE Transactions on Systems, Man and Cybernetics, 17(2):187–198, 1987.

[12] F. C. García-López, M. García-Torres, B. Melián-Batista, J. A. M. Pérez, and J. M. Moreno-Vega.
Solving the feature selection problem by a parallel scatter search.
European Journal of Operations Research, 169(2):477–489, 2006.

[13] I. Guyon and A. Elisseeff.
An introduction to variable and feature selection.
Journal of Machine Learning Research, 3:1157–1182, 2003.

[14] M. Ichino and J. Sklansky.
Feature selection for linear classifier.
In *Proceedings of the Seventh International Conference on Pattern Recognition*, pages 124–127. Morgan Kaufmann, 1984.

[15] M. Ichino and J. Sklansky.
Optimun feature selection by zero-one programming.
IEEE Transactions on Systems, Man and Cybernetics, 14(5):737–746, 1984.

[16] I. Inza, P. Larrañaga, R. Etxeberria, and B. Sierra.
Feature subset selection by bayesian networks based optimization.
Artificial Intelligence, 123(1-2):157–184, 2000.

[17] J. M. Keynes.
A treatise on probability.
Mcmillan and Co., 1921.

[18] K. Kira and L. A. Rendell.
The feature selection problem.
In *In Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 129–134. Menlo Park: AAAI Press / The MIT Press, 1992.

CONACYT

[19] R. Kohavi and G. H. John.
Wrappers for feature subset selection.
Artificial Intelligence, 97(1-2):273–324, 1997.

[20] D. Koller and M. Sahami.
Toward optimal feature selection.
In *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 284–292, 1996.

[21] D. Koller and M. Sahami.
Toward optimal feature selection.
In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 284–292, 1996.

[22] I. Kononenco.
Estimating attributes: Analysis and extension of relief.
In *In Proceedings of the European Conference on Machine Learning*, pages 171–182. Springer-Verlag, 1994.

[23] P. Langley and S. Sage.
Oblivious decision trees and abstract cases.
In *Working Notes of the AAAI-94 Workshop on Case-Based Reasoning*. AAAI Press, 1994.

[24] H. Liu, H. Motoda, and M. Dash.
A monotonic measure for optimal feature slection.
In C. Nedellec and C. Rouveirol, editors, *Machine Learning: ECML-98*, pages 101–106. Springer-Verlag, 1998.

[25] H. Liu and R. Setiono.
Feature selection and classification - a probabilistic wrapper approach.
In *Proceedings of the Ninth International Conference on Industrial and Engineering Applications of AI and ES*, pages 419–424, 1996.

CONACYT

[26] H. Liu and R. Setiono.
A probabilistic approach to feature selection - a filter solution.
In *Proceedings of International on Machine Learning*, pages 319–327. Morgan Kaufmann, 1996.

[27] H. Liu and R. Setiono.
Scalable feature selection for large sized databases.
In *Proceedings of the Fourth World Congress on Expert Systems*, volume 1. Morgan Kaufmann, 1998.

[28] M. Modrzejewski.
Feature selection using rough sets theory.
In *Proceedings of the European Conference on Machine Learning*, pages 213–226, 1993.

[29] A. W. Moore and M. S. Lee.
Efficient algorithms for minimizing cross validation error.
In *International Conference on Machine Learning*, pages 190–198, 1994.

CONACYT

[30] A. N. Mucciardi and E. E. Gose.
A comparison of seven techniques for choosing subsets of pattern recognition.
IEEE Transactions on Computers, 20:1023–1031, 1971.

[31] P. M. Narenda and K. Fukunaga.
A branch and bound algorithm for feature subset selection.
IEEE Trans. on Computer, 26(9):917–922, 1977.

[32] A. L. Oliveira and A. S. Vincentelli.
Constructive induction using a non-greedy strategy for feature selection.
In *Proceedings of the Ninth International Conference on Machine Learning*, pages 355–360. Morgan Kaufmann, 1992.

[33] P. Pudil, J.Novovicova, and J. Kittler.
Floating search methods in feature selection.
Pattern Recognition Letters, 15:1119–1125, 1994.

CONACYT

[34] M. Scherf and W. Brauer.
Improving rbf networks by the feature selection approach eubafes.
In *Proceedings of the 7th International Conference on Artificial Neural Networks*, pages 391–396, 1997.

[35] J. C. Schlimmer.
Efficiently inducing determinations: A complete and systematic search algorithm that uses optimal pruning.
In *In Proceedings og the Tenth Inetrnational Conference on Machine Learning*, pages 284–290. ICML93, 1993.

[36] J. Segen.
Feature selection and constructive inference.
In *In Proceedings of the Seventh International Conference on Pattern Recognition*, pages 1344–1346, 1984.

[37] J. Sheinvald, B. Dom, and W. Niblack.
A modelling approach to feature selection.
In *In Proceedings of the Tenth International Conference on Pattern Recognition*, pages 535–539, 1990.

[38] W. Siedlecki and J. Sklansky.
A note on genetic algorithms for large-scale feature selection.
Pattern Recognition Letters, 10:335–347, 1989.

[39] H. Vafaie and I. F. Imam.
Feature selection methods: Genetic algorithms vs. greedy-like search.
In *Proceedings of the International Conference on Fuzzy and Intelligent Control Systems*, 1994.

CONACYT

[40] H. Wang, D. Bell, and F. Murtagh.
Axiomatic approach to feature subset selection based on relevance.
IEEE Transactions on Pattern Analysis and machine Intelligence, 21(3):271–277, 1999.

[41] K. Wang and S. Sundaresh.
Selecting features by vertical compactness of data.
In D. Heckerman, H. Mannila, and D. Pregibon, editors, *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pages 275–278. AAAI Press, 1997.

[42] L. Xu, P. Yan, and T. Chang.
Best first strategy for feature selection.
In *In Proceedings of the Ninth International Conference on Pattern Recognition*, pages 706–708, 1988.

CONACYT

[43] L. Yu and H. Liu.
Efficient feature selection via analysis of relevance and redundancy.
J. Mach. Learn. Res., 5:1205–1224, 2004.

CONACYT