Feature Selection

Jeonghun Yoon

Feature

Feature, variable, attribute, demension

- variable : the **raw** input variables
- features : **variables constructed** for the input variables

Objective of feature selection

목적 : 좋은 모델(또는 예측기good predictor)를 생성하기 위하여 유용한useful feature들의 subset을 선택

- Relevant variable? 만약에 그 variable이 redundant 하다면?
- Feature의 집합 $F = \{f_1, \dots, f_i, \dots, f_n\}$이 주어졌을 때, 학습 모델이 pattern을 잘 분류해낼 수 있도록 학습 모델의 성능 및 능력을 극대화할 수 있는 feature의 subset $F' \subseteq F$를 찾는다.
- 모델을 간단하게 만들어 사용자가 모델을 해석하기 쉽도록 한다.
- 학습 알고리즘의 성능을 향상시키며 학습 시간을 줄인다.
- Overfitting을 제한하며 모델의 일반화를 더 강화한다.
- Interpretability
- Gain a deeper insight into the underlying processes that generates the data

When feature selection is important

- Noise data
- Lots of low frequent features
- Use multi-type features
- Too many features comparing to samples
- Complex model
- Samples in real scenario is inhomogeneous with training & test samples

머신러닝에서의 Feature selection

- 분류기의 경우 taget class, 회귀의 경우 target에 대한 정보는 input variables 에서 고유하게 결정된다. 즉 variables은 target에 대한 정보의 본질이라고 볼 수 있다.
- 더 많은 정보를 가지고 있다고 해서 모델의 결정력discrimination power가 증가하는 것은 아니다.
- Feature의 dimensionality와 모델의 performance의 관계
  - 차원의 저주에 의하여, variable의 수가 증가할수록, 좋은 성능을 가지는 모델을 학습하기 위한 샘플의 수는 지수적으로exponentially 증가한다.
  - 분류기의 성능은 feature의 수가 매우 클 때, 감소한다.

Characteristics feature (subset) selection

Quality of the feature subset is very essential to determine performance of algorithm. The number of data needed to be processed depends on the dimensionality of features.

- There are much irrelevant, redundant information present or noisy with unreliable data, then knowledge discovery during the training phase is merely waste.
- In real-world data, the representation of data often consists of too many features, but only a few of them might be related to the target concept.
- Redundancy is always expected and that should be carefully handled and removed.

Characteristics feature (subset) selection

Relevant

- These features have an influence on the output and its role can‟t be assumed by the rest.

Irrelevant

- These features do not have any influence on the output and it has values random for each instances.

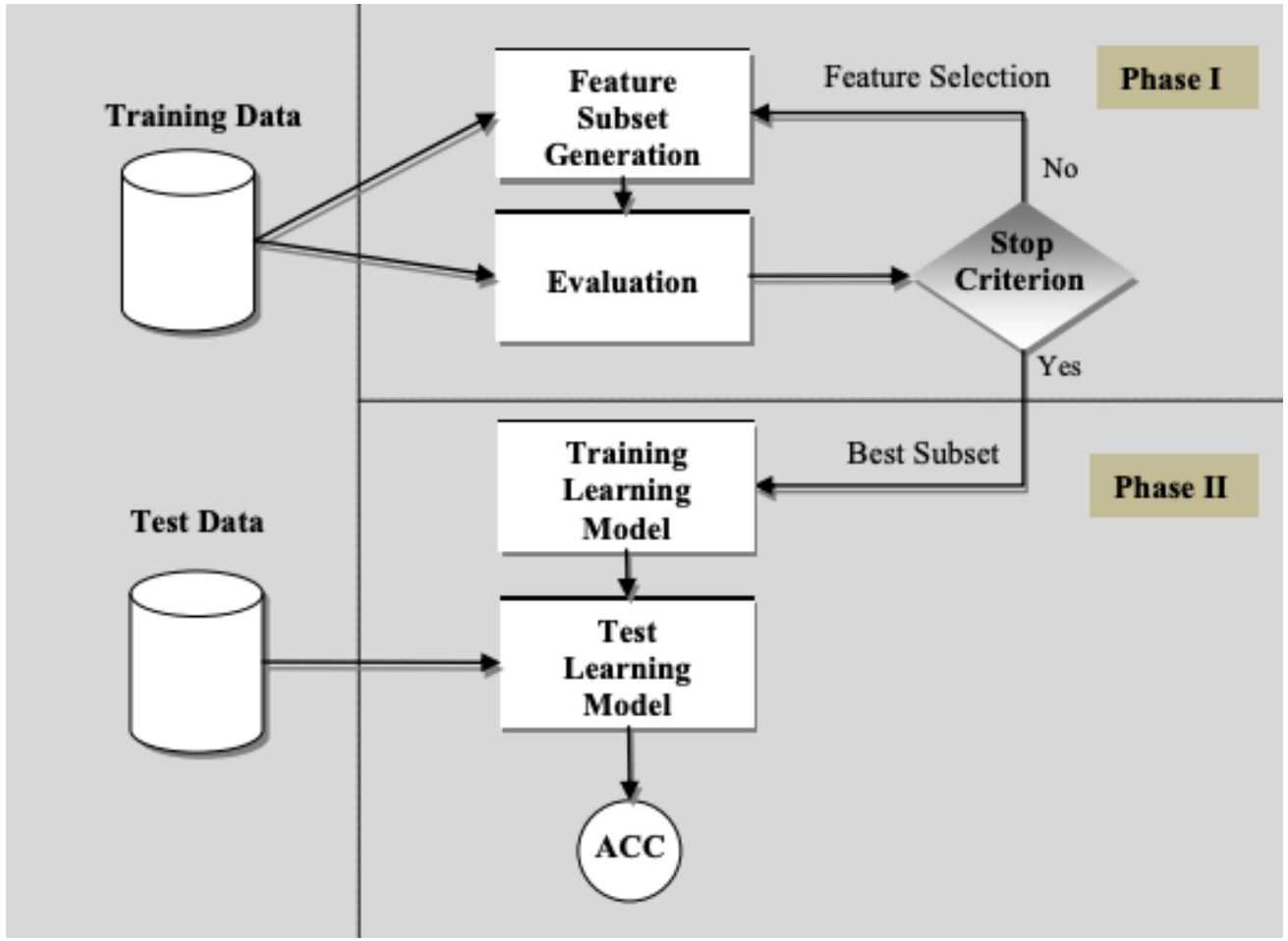
Redundant:

- A redundancy exists whenever a feature takes the role of another feature. Another notion of redundancy is correlation.

Useful? Relevant?

Selecting subsets of features that are useful to build a good predictor. This contrasts with the problem of finding or ranking all potentially relevant variables. Selecting the most relevant variables is usually suboptimal for building a predictor, particularly if the variables are redundant. Conversely, a subset of useful variables may exclude many redundant, but relevant, variables.

Feature (subset) selection process



Feature (subset) selection process

A typical feature selection process has two phases:

- Feature selection
- Model fitting performance evaluation

The feature selection phase further contains three steps:

- Generating a candidate subset from the original features via certain research strategies
- Evaluating the candidate set by estimating the utility of the features in the candidate set. Based on the evaluation, some features in the candidate set might be discarded or new ones might be added to the selected feature set according to their relevance.
- Determining the current set of selected features are good enough using certain stopping criterion, i.e. a feature selection methods returns the selected set of features, else iterates until the stopping criterion is met.

Feature (subset) selection process

Starting Point

- The search for feature subsets starts with no features or with the full features. The first approach is forward selection whereas the second approach is backward selection.

Search Strategy

- The best subset of features could be found by evaluating all the possible subsets, which is known as exhaustive search.
- This exhaustive search needs to search all of 2^n possible subsets of n features, so it becomes impractical with huge volume of features.
    - More realistic practical approach needs to be considered.

Feature (subset) selection process

Subset Evaluation

Based on these two approaches, the goodness (ie: classification accuracy) of executing algorithm among the preprocessed dataset is determined.

- Filter approach
- wrapper approach

Stopping Criteria

It is necessary to halt the search once the number of selected features reached a predetermined threshold value.

- Reach the end and the search gets completed.
- Reach the specified boundary which is set externally (minimum number of features or maximum number of iterations).
- Find out reasonable good number of good subset

Feature (subset) selection process

Result Validation

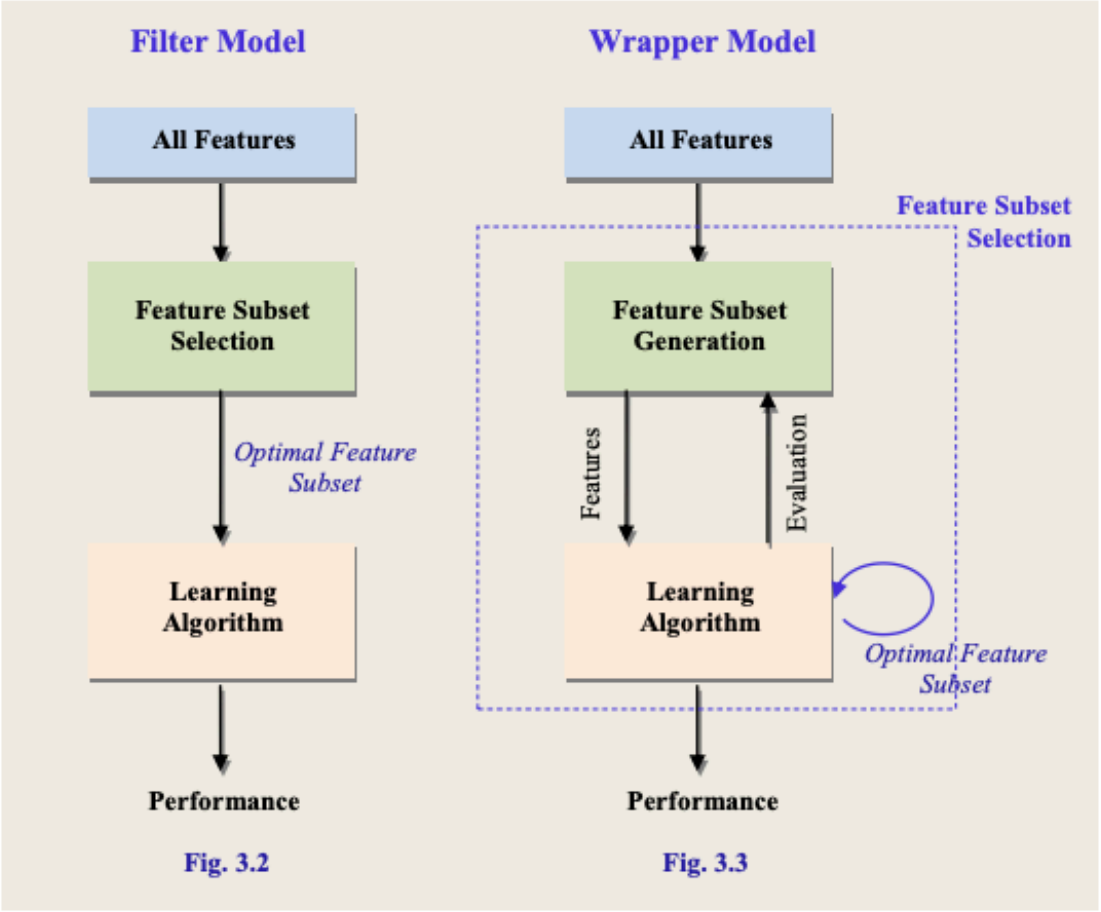
- Some indirect methods are used by monitoring the change of mining performance with the change of features
- For example, classification error rate is used as a performance indicator for a learning task, for a selected feature subset, perform “before-after” experiment.

Search Strategy

Depends on how the selected features are evaluated, different strategies could be adopted, which broadly fall into three categories:

- Filter model
- Wrapper model
- Embedded model : utilizing the filter model inside the wrapper model

Search Strategy



Subset Evaluation

Filter model

- Relied on analyzing the general characteristics of data evaluating features without involving any learning algorithms.

Wrapper model

- Require a predetermined learning algorithm

Embedded model

- Incorporated feature selection as a part of the model fitting / training process, the features utility is obtained based on analyzing their utility for optimizing the objective function of the learning model.

Advantages of Filter model

- Independent of any learning model and is free from bias associated with any learning models.
- It allows the algorithms to have very simple structure, which usually employed as straightforward search strategy, such as backward elimination or forward selection.
- The filter model is easy to design and also easy to understand : most features selection algorithms uses the filter model.

An Introduction to Variable and Feature Selection

- http://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf

An Introduction to Variable and Feature Selection

Isabelle Guyon ISABELLE@CLOPINET.COM
Clopinet
955 Creston Road
Berkeley, CA 94708-1501, USA

André Elisseeff ANDRE@TUEBINGEN.MPG.DE
Empirical Inference for Machine Learning and Perception Department
Max Planck Institute for Biological Cybernetics
Spemannstrasse 38
72076 Tübingen, Germany

Editor: Leslie Pack Kaelbling

Abstract

Variable and feature selection have become the focus of much research in areas of application for which datasets with tens or hundreds of thousands of variables are available. These areas include text processing of internet documents, gene expression array analysis, and combinatorial chemistry. The objective of variable selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data. The contributions of this special issue cover a wide range of aspects of such problems: providing a better definition of the objective function, feature construction, feature ranking, multivariate feature selection, efficient search methods, and feature validity assessment methods.

Keywords: Variable selection, feature selection, space dimensionality reduction, pattern discovery, filters, wrappers, clustering, information theory, support vector machines, model selection, statistical testing, bioinformatics, computational biology, gene expression, microarray, genomics, proteomics, QSAR, text classification, information retrieval.

(Variable ranking)

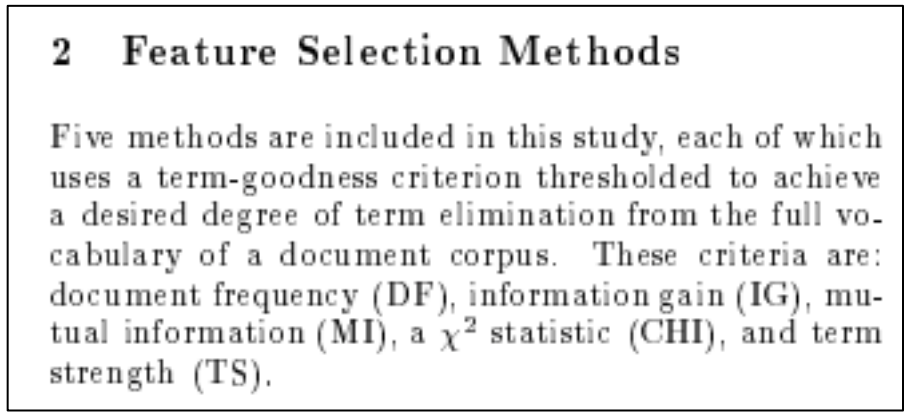
1. Correlation Criteria

2. Single Variable Classifiers

3. Information Theoretic Ranking Criteria

A comparative study on feature selection in text categorization

- http://courses.ischool.berkeley.edu/i256/f06/papers/yang97comparative.pdf



2 Feature Selection Methods

Five methods are included in this study, each of which uses a term-goodness criterion thresholded to achieve a desired degree of term elimination from the full vocabulary of a document corpus. These criteria are: document frequency (DF), information gain (IG), mutual information (MI), a χ^2 statistic (CHI), and term strength (TS).

1. Document frequency
2. Information gain
3. Mutual information
4. χ^2 statistic
5. Term stength

Feature selection for text classification Based on Gini Coefficient of Inequality

- http://proceedings.mlr.press/v10/sanasam10a/sanasam10a.pdf

Feature Selection for Text Classification Based on Gini Coefficient of Inequality

Sanasam Ranbir Singh RANBIR@IITG.ERNET.IN
Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
Guwahati 781039, Assam, India

Hema A. Murthy HEMA@LANTANA.TENET.RES.IN
Department of Computer Science and Engineering
Indian Institute of Technology Madras
Channai 600036, Tamil Nadu, India.

Timothy A. Gonsalves TAG@LANTANA.TENET.RES.IN
Department of Computer Science and Engineering
Indian Institute of Technology Madras
Channai 600036, Tamil Nadu, India

Editor: Huan Liu, Hiroshi Motoda, Rudy Setiono, and Zheng Zhao

Abstract

A number of feature selection mechanisms have been explored in text categorization, among which mutual information, information gain and chi-square are considered most effective. In this paper, we study another method known as *within class popularity* to deal with feature selection based on the concept *Gini coefficient of inequality* (a commonly used measure of inequality of *income*). The proposed measure explores the relative distribution of a feature among different classes. From extensive experiments with four text classifiers over three datasets of different levels of heterogeneity, we observe that the proposed measure outperforms the mutual information, information gain and chi-square static with an average improvement of approximately 28.5%, 19% and 9.2% respectively.

Keywords: Text categorization, feature selection, gini coefficient, within class popularity

1. Gini Coefficient of Inequality

Variable ranking

- Frequency based
- Correlation Criteria
- Single Variable Classifiers
- Gini indexing
- Information gain
- Mutual information
- Chi square statistic
- MRMR (Maximum Relevance-Minimum Redundancy)

Correlation Criteria

- The Pearson correlation coefficient is defined as:

$$\mathcal{R}(i) = \frac{cov(X_i, Y)}{\sqrt{var(X_i)var(Y)}}$$

- The estimator of $R(i)$

$$R(i) = \frac{\sum_{k=1}^{m}\left(x_{k,i} - \bar{x_i}\right)(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{m}\left(x_{k,i} - \bar{x_i}\right)^2 \sum_{k=1}^{m}(y_k - \bar{y})^2}}$$

- Correlation criteria such as $R(i)$ can only detect linear dependencies between variable and target.
- 높은 상관관계를 가질 수록 높은 score를 가지게 된다.
- $R(x_i, y) \in [-1, 1]$이기 때문에, 주로 $R(x_i, y)^2$ 또는 $|R(x_i, y)|$를 사용한다.

Single Variable Classifiers

- In the classification case the idea of selecting variables according to their individual predictive power, using as criterion the performance of a classifier built with a single variable.
  - The value of the variable itself can be used as discriminant function.
  - A classifier is obtained by setting a threshold θ on the value of the variable.
- The predictive power of the variable can be measured in terms of error rate or FPR or FNR.

Gini Indexing

- The Gini coefficient (also known as the Gini index or Gini ratio) is a measure of statistical dispersion intended to represent the income distribution.
- It measures the inequality among values of a frequency distribution.
- The Gini coefficient could theoretically range from 0 (complete equality) to 1 (complete inequality).

$$G = \left| 1 - \sum_{k=1}^{n} (X_k - X_{k-1})(Y_k + Y_{k-1}) \right|$$

- G : Gini coefficient
- X_k : cumulated proportion of the one variable for $k = 0, \dots, n$ with $X_0 = 0, X_n = 1$
- Y_k : cumulated proportion of the one variable for $k = 0, \dots, n$ with $Y_0 = 0, Y_n = 1$

Information Gain

- Due to its computational efficiency and simple interpretation, information gain is one of the most popular feature selection methods. It is used to measure the dependence between features and labels.

- Information gain between the i-th feature f_i and the class labels C :

$$IG(f_i, C) = H(f_i) - H(f_i|C)$$

- $H(f_i)$ is the entropy of f_i and $H(f_i|C)$ is the entropy of f_i after observing C :
  - $H(f_i) = -\sum_j p(x_j) \log_2(p(x_j))$
  - $H(f_i|C) = -\sum_k p(c_k) \sum_j p(x_j|c_k) \log_2(p(x_j|c_k))$

- In information gain, a feature is relevant if it has a high information gain.

Mutual information

- 연속형 feature

$$I(x_i, y) = \int_{x_i} \int_{y} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)p(y)} dxdy$$

- 이산형 feature

$$I(x_i, y) = \sum_{x_i} \sum_{y} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)p(y)} dxdy$$

- For text classification,
  - Two away contingency table of a term t and a category C_1

	C_1	Not C_1
Term t	A	B
Not term t	C	

The total number of documents

$$MI(t, C_1) = \log \frac{p(t \cap C_1)}{p(t)p(C_1)} \approx \log \frac{A \times N}{(A + C) \times (A + B)}$$

  - $MI(t, C_1)$ has a natural value of zero if t and C_i are independent.

χ^2 Statistic

- Chi-Square is the common statistical test that measures the divergence from the distribution expected if the assumed feature occurrence is actually independent of the class value.
- When the Chi-square Statistics is larger than the critical value determined by the degrees of freedom, then the feature and the class are considered dependent and such features are selected.

$$\chi^2(f) = \sum_{v \in V} \sum_{i=1}^{m} \frac{\left(A_i(f = v) - E_i(f = v)\right)^2}{E_i(f = v)}$$

- $A_i(f = v)$: the number of instances in class c_i with $f = v$
- $E_i(f = v)$: the expected value of $A_i(f = v)$ calculated as $p(f = v)p(c_i)N$

χ^2 Statistic

- This method measures the lack of independence between the terms of category.
- In statistics, the χ^2 test is applied to test the independence of two events.
  - The two events A and B are defined to be independent if $p(AB) = p(A)p(B)$ or equivalently, $p(A|B) = p(A)$, $p(B|A) = P(B)$

Maximum Relevance-Minimum Redundancy (MRMR)

- MRMR is a scheme in feature selection to adopt the features that correlate the strongest correlation with a classification variable.
- The MRMR methods use the mutual information between a feature and a class as relevance of the feature for the class.

Maximum Relevance-Minimum Redundancy (MRMR)

- Maximal Relevance is to search feature set S satisfying:

$$\max D(S,c) \quad \text{where} \quad D(S,c) = \frac{1}{|S|}\sum_{x_i \in S} I(x_i, c)$$

- $I(x_i, c)$: the mutual information between feature x_i and class c

- Minimal Redundancy is to search feature set S satisfying:

$$\min R(S) \quad \text{where} \quad R(S) = \frac{1}{|S|^2}\sum_{x_i, x_j \in S} I\left(x_i, x_j\right)$$

- $I\left(x_i, x_j\right)$: the mutual information between feature x_i and x_j

- Maximum Relevance-Minimum Redundancy

$$\max \phi(D,R) \quad \text{where} \quad \phi(D,R) = D - R$$

Questions

- Can Presumably Redundant Variables Help Each Other?

  One common criticism of variable ranking is that it leads to the selection of a redundant subset. The same performance could possibly be achieved with a smaller subset of complementary variables.

  - Noise reduction and consequently better class separation may be obtained by adding variables that are presumably redundant.

- How Does Correlation Impact Variable Redundancy?
  - Perfectly correlated variables are truly redundant in the sense that no additional information is gained by adding them.
  - Very high variable correlation (or anti-correlation) does not mean absence of variable complementarity.

Questions

- Can a Variable that is Useless by Itself be Useful with Others?
  - A variable that is completely useless by itself can provide a significant performance improvement when taken with others.
  - Two variables that are useless by themselves can be useful together.

References

- https://shodhganga.inflibnet.ac.in/bitstream/10603/137885/3/12%20chapter%203.pdf
- https://www.ic.unicamp.br/~wainer/cursos/1s2012/mc906/FeatureSelection.pdf
- https://pdfs.semanticscholar.org/310e/a531640728702fce6c743c1dd680a23d2ef4.pdf