Feature Selection in Unsupervised Way Using Feature Similarity

- Titli Sarkar
- txs7980@louisiana.edu

Why feature selection?

- Mining dataset of huge dimension is a problem !
- Dataset can be represented by a subset of original features
  - should preserve principal characteristics of the whole dataset
- Feature selection and Dimensionality Reduction

Conventional Methods

- Evaluate different feature subsets using an *index* and select best among them.
  - index defines the capability of respective subsets in classification (supervised) or clustering (unsupervised)
- Problem !
  - Significantly high 'searching' time in high-dimensional dataset
    - several heuristic search techniques are adopted:
      - Branch and Bound algorithm, suggested by Devijver and Kitter [1]
      - Greedy algorithms like Sequential Forward and Backward Search [1]
      - Robust methods like GA (Genetic Algorithm) [3]
  - Feature engineering is not possible if domain knowledge is not available, so, using unsupervised way of learning is preferable

Unsupervised Methods

Method 1: maximize clustering performance, quantified by some *index*

- Sequential Unsupervised Feature Selection Algorithm
- Wrapper Approach based on EM
- Maximum Entropy based method
- Neuro-fuzzy approach (newly developed)

Method2: selection of features based on feature dependency and relevance (Our Concern)

- Redundant features carrying little or no information are removed
- Dependent Measures:
  - Correlation Coefficients
  - Measure of Stastical Dependency
  - Linear Dependence

Our Approach

- Use an unsupervised algorithm
- Use feature dependency/similarity for redundancy reduction
  - Similarity measure/ index : maximal information compression index
- Partition the feature set into distinct homogeneous subsets or clusters
- Choose one representative feature from each cluster and add them to a new feature set
- Represent the dataset with new feature set
- Evaluate the quality of clustering
- Get rid of ‘searching’!

Dataset

- TSR-keys for a set of proteins (Dr. Sumi Singh, Dr. Xu Wu, Dr. Vijay Raghavan)
- Keys are the representative features.
- We need to represent the documents as vectors of representative features.

Unsupervised feature selection Approach

- Step1: Partition the feature set in clusters
- Step 2: Choose one representative features from each cluster, discard other features in the corresponding clusters.
- Partitioning is based on K-NN (k-Nearest Neighbors) principle using *maximal information compression index* feature similarity measure.
- First, compute k-nearest features of each feature.
- Keep the features having the most compact subset, discard other k neighboring features
  - most compact subset of feature is selected as determined by its distance from farthest neighbor

Choice of ‘k’

- k controls the size of the reduced set, initially set to <= D-1 where D is the dimension of the dull dataset
- a constant error-threshold Ɛ is introduced in measuring the k-nearest neighbors
- Ɛ set to distance of the k-th nearest neighbor in first iteration
- In later equations, compare value of λ_2 (eigenvalue), corresponding to each subset of feature with Ɛ,
  - - if λ_2 > Ɛ , decrease value of k.
- *k* determines the error threshold *(Ɛ),* the representation of the data at different degrees of details is controlled by its choice. This characteristic is useful in data mining where *multiscale* representation of the data is often necessary.

Similarity Measure

- Maximal Information Compression Index (λ_2):

Let ∑ be the covariance matrix of random variables x and y. Define, maximal information compression index as λ_2 (x, y) = smallest eigenvalue of ∑, i.e.,

$$2\lambda_2(x, y) = (\text{var}(x) + \text{var}(y) - \sqrt{(4\text{var}(x)\text{var}(y)\,(1 - \rho(x, y)^2))}$$

The value of λ_2 is zero when the features are linearly dependent and increases as the amount of dependency decreases. It may be noted that the measure λ_2 is nothing but the eigenvalue for the direction normal to the principle component direction of feature pair (x, y).

λ_2 is the amount of reconstruction error committed if the data is projected to a reduced dimension in the best possible way. Therefore, it is a measure of the *minimum amount of information loss* or the *maximum amount of information compression,* possible.

Feature selection Method[2]

Step 1: Choose an initial value of $k < D - 1$. Initialize the reduced feature subset R to the original feature set 0, i.e., $R = 0$.

Step 2: For each feature $F_i \in R$, compute r^k_i .

Step 3: Find feature F_i, for which r^k_i is minimum.

Retain this feature in R and *discard* k nearest features of $F_{i'}$. (Note: $F_{i'}$ denotes the feature for which removing k nearest-neighbors will cause minimum error among all the features in R). **Let** $\epsilon = r^k_i$.

Step 4: If $k >$ cardinality(R) — 1: $k =$ cardinality(R) — 1.
Step 5: If $k = 1$: **Go to Step 8.**

Step 6: While $r^k_{i'} > \epsilon$ **do:**

a) $k = k - 1$.

$r^k_{i'} = \inf F_i \in R\ r^k_i$

("k" is decremented by 1, until the "k-th nearest-neighbor" of at least one of the features in R is less than e-dissimilar with the feature)

If $k = 1$: **Go to Step 8.**

(if no feature in R has less than ϵ-dissimilar "nearest-neighbor" select all the remaining features in R)

End While

Step 7: Go to Step 2.

Step 8: Return feature set R as the reduced feature set.

Feature evaluation

- need some index for evaluating the effectiveness of the selected feature subsets
- Following two category of methods can be considered as evaluation measure:
  1. *need class information of the samples:*
     - class separability
     - K-NN classification accuracy
     - naive Bayes classification accuracy
  2. *do not need class information of the samples:*
     - Entropy
     - fuzzy feature evaluation index
     - representation entropy

Conclusion

- This unsupervised method of is expected to handle the challenge of feature selection on very high dimensional data in Data Mining area

- The weakness of other feature reduction methods in high dimension is time involved in searching. Absence of searching in this method promises better computational time.

- Its difficult to use any supervised algorithm when domain knowledge is not available. Unsupervised feature selection using feature similarity can best fit in this context.

References

1. P.A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach.* Englewood Cliffs: Prentice Hall, 1982.
2. P. Mitra, C. A. Murthy and S. K. Pal. Unsupervised Feature Selection using Feature Similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24(3), pages 301-312, 2002.