A Graph-based Approach to Feature Selection

Zhihong Zhang and Edwin R. Hancock
Department of Computer Science
The University of York
YO10 5GH, UK

The idea underpinning feature selection

- Reduce the dimensionality of the feature space
- Speed up and reduce the cost of a learning algorithm
- Obtain the feature subset which is most relevant but less redundant to classification

The existing methods

- Variance based methods
  - PCA based feature selection *[S.Buchala, N.Davey , T.M. Gale and R.J. Frank, 2005]*
  - Limitation: only consider the variance of features, nothing to do with the classification

- Mutual information based methods
  - MIFS: mutual information feature selection *[R.Battiti, 2002*
  - MRMR: maximum-relevance minimum redundancy *[H.Peng, 2005]*
  - JMI: joint mutual information *[H.Yang, 1999]*
  - Limitation: based on the assumption that either that features independently influence the class variable or do so only involving pair-wise feature interaction

Graph-based Refinement of feature-set

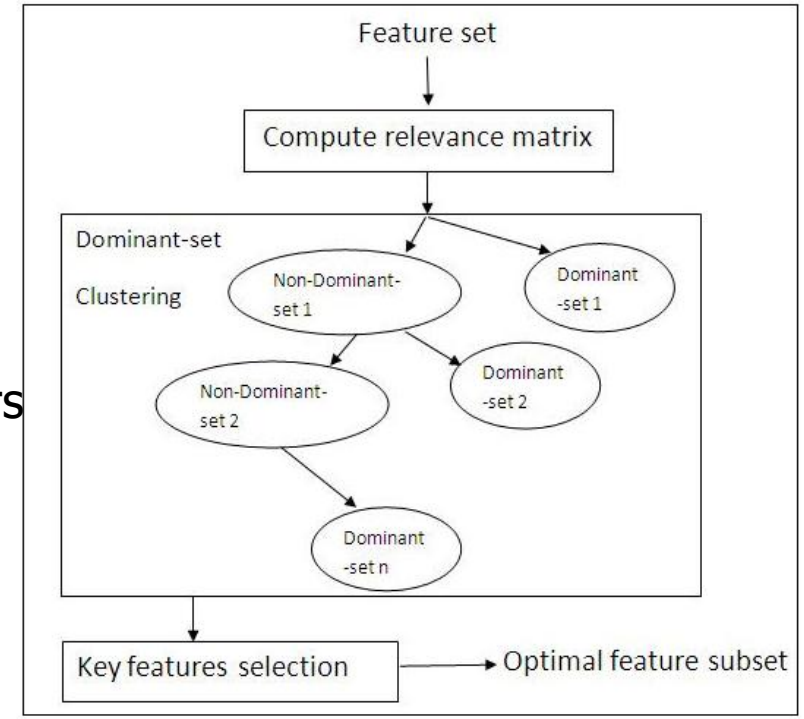
- Characterise relevance of feature vectors using graph-based representation of mutual information.
- Cluster feature vectors F into dominant sets using mutual information
- Select optimal feature subsets f from each dominant set using multidimensional interaction information.

Our method: Dominant-set clustering & Multidimensional interaction information

- First constructs a graph in which each node corresponds to each feature, and each edge has a weight corresponding to the interaction information among features connected by that edge.
- Then perform dominant-set clustering to select a highly coherent set of features using pair-wise similarities
  - Advantage: Separates features into clusters prior to selection, therby allowing us to limit the search space for higher order interactions
- Finally selects features based on a new measure called the multidimensional interaction information (MII)
  - Advantage: it is capable of detecting the relationships between third or higher order features combinations

The flowchart of our approach for feature selection

- Using the graph representation of the features, there are three steps to the algorithm, namely
  - Computing the relevance matrix based mutual information between feature vectors
  - Dominant-set clustering to cluster the feature vectors
  - Applying MII criterion into each dominant-set to rank the features and then select the top k key features based on the value of incremental gain

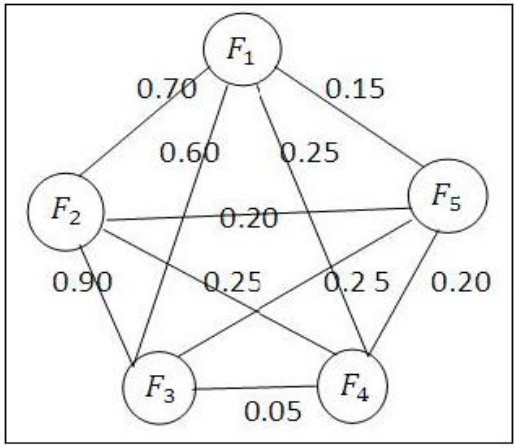


Clustering feature-vectors

Find best set for feature selection based on mutual information criterion.

The concept of Dominant-set clustering

- Dominant set *[M.Pavan, M.Pelillo, , 2003]*
  - Definition: The dominant set, is a combinational concept in graph theory that generalizes the notion of a maximal complete subgraph from simple graphs to edge-weighted graphs. In fact, dominant sets turn out to be equivalent to maximal cliques.
  - The definition of the dominant set simultaneously emphasizes internal homogeneity and together with external inhomogenety. Thus it is can be used as a general definition of a "cluster".
  - Example: features {F1, F2, F3}form the dominant set, since the edge weights ``internal" to that set (0.6, 0.7 and 0.9) are larger than the sum of those between the internal and external features (which is between 0.05 and 0.25).



Mutual Information

- Shannon entropy

$$H(Y) = \sum_{y \in Y} P(y) \log P(y)$$

- Conditional entropy

$$H(Y \mid X) = -\int p(x) \left\{ \sum_{y \in Y} p(y \mid x) \log p(y \mid x) \right\} dx$$

- Mutual information

$$I(X : Y) = H(Y) - H(Y \mid X)$$

$$= \sum_{y \in Y} \int p(y \mid x) \log \frac{p(y, x)}{p(y)\, p(x)} dx$$

Elements of weight matrix

- Measure joint relevance of feature vectors using mutual information

$$W(u,v) = \frac{I(F_u, F_v)}{H(F_u) + H(F_v)}$$

- Dominant sets selects largest set of most relevant (least redundant) features.

Locate the dominant-set

- Given the graph G=(V,E), we can locate the dominant set by finding the solutions of a quadratic program that maximizes the functional

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{W}\mathbf{x} \ . \qquad (1)$$

subject to $\mathbf{x} \in \triangle$, where $\triangle = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq 0 and \sum_{i=1}^{n} x_i = 1\}$ and W is the relevance weight matrix between features.

- We can get the solution of above equation using a iterative update equation:

$$x_i(t+1) = x_i(t)\frac{(\mathbf{W}\mathbf{x}(t))_i}{\mathbf{x}(t)^T\mathbf{W}\mathbf{x}(t)} \ . \qquad (2)$$

where $x_i(t)$ is correspondent to the i-th feature vector at iteration t of the update process.

Dominant-set clustering

- We can formulate the dominant-set clustering algorithm in the following:

Input: the similarity matrix $\mathbf{W}$

1. Initialize $\mathbf{W}^k, k = 1$ with $\mathbf{W}$
2. Calculate the local solution of (1) by (2): $\boldsymbol{u}^k \; and \; f(\boldsymbol{u}^k)$
3. Get the dominant set: $\boldsymbol{S}^k = \sigma(\boldsymbol{u}^k)$
4. Split out $\boldsymbol{S}^k$ from $\mathbf{W}^k$ and get a new affinity matrix $\mathbf{W}^{k+1}$
5. If $\mathbf{W}^{k+1}$ is not empty, $\mathbf{W}^k = \mathbf{W}^{k+1}$ and $k = k + 1$, then go to step 2; else exit

Output: $\cup_{l=1}^{k}\{\boldsymbol{S}^l, \boldsymbol{u}^l, f(\boldsymbol{u}^l)\}$

Feature-component selection

Select components of feature-vectors based on multidimensional interaction information

Selecting key features

- The multidimensional interaction information between feature vector $F = \{f_1, \ldots, f_m\}$ and class variable C is:

$$I(F;C) = I(f_1, \ldots, f_m; C) = \sum_{f_1, \ldots, f_m} \sum_{c \in C} P(f_1, \ldots, f_m; c) \times \log \frac{P(f_1, \ldots, f_m; c)}{P(f_1, \ldots, f_m) P(c)} .$$

- Using Parzen windows for probability distribution estimation, we then apply the greedy strategy to select the feature that maximizes the multidimensional mutual information between the features and the output class set. As a result the first feature f'_{max} maximizes $I(f', C)$, the second selected feature f''_{max} maximizes $I(f'', f', C)$ and so on. For each dominant set, we repeat this procedure to rank the features and meanwhile record the incremental gain for each feature.

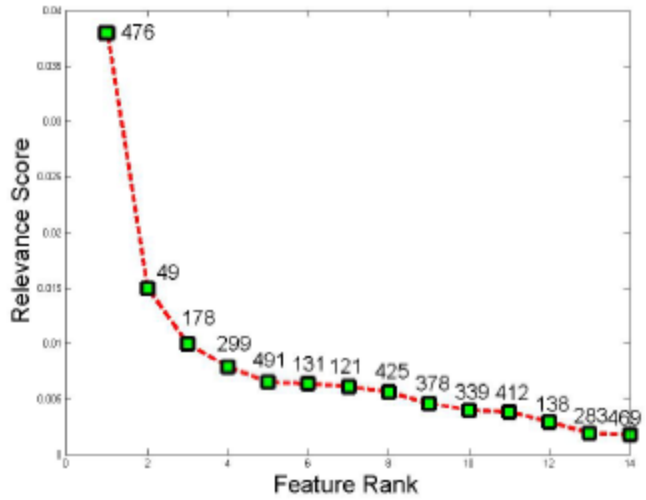
Experiments

- Benchmark data sets

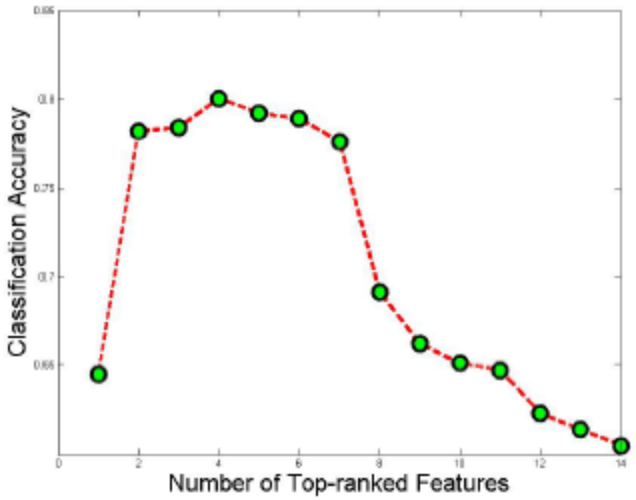
Data-set	From	Examples	Features	Classes
Madelon	NIPS	2000	500	2
Breast cancer	UCI	699	10	2
Pima	UCI	768	8	2
Australian	UCI	690	14	2

Madelon data set using MRMR algorithm

- The result on Madelon data set using MRMR for feature ranking. The values of the relevance score for the top 14 features are presented in the left part along with the feature indices, while the classification accuracies are plotted in the right part .



(a) Top-ranked Features by Relevance Score

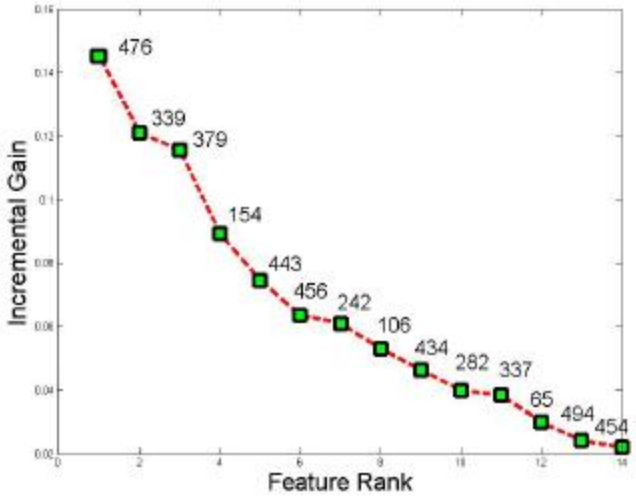


(b) Classification result

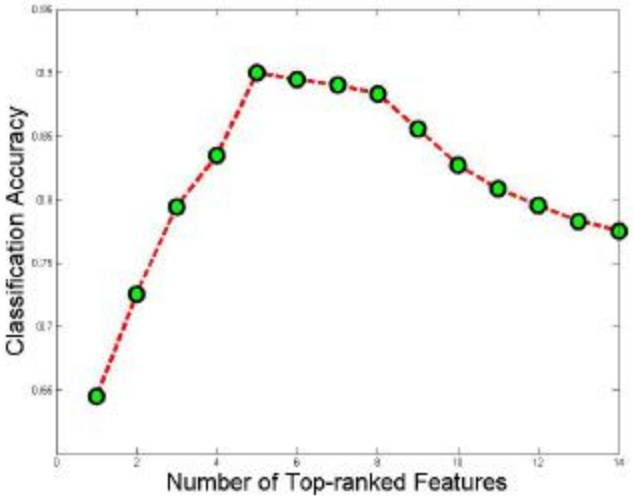
- Best accuracy 80.5% using 4 features

Madelon data set using our algorithm

- The result on Madelon data set for our algorithm. The values of the Incremental gain for the top 14 features are presented in the left part along with the feature indices, while the classification accuracies are plotted in the right part



(a) Top-ranked Features by Incremental Gain



(b) Classification result

- We achieve accuracy 90% using the leading 5 features.

- The classification accuracy on the top features selected by different methods in the Breast Cancer data set

No.of Features Selected	DS*plus*MII	k-means	MRMR	MIFS
2	88.84%	78.97%	88.26%	88.26%
4	96.3%	78.97%	87.55%	82.51%
6	96.3%	91.55%	87.55%	82.51%

- The classification accuracy on the top features selected by different methods in the Australian data set

No.of Features Selected	DS*plus*MII	k-means	MRMR	MIFS
4	85.51%	62.32%	58.11%	58.26%
8	74.78%	66.37%	58.11%	58.11%
12	74.78%	66.23%	74.78%	62.75%

Conclusion

- We have presented a new graph theoretic approach to feature selection.
- Dominant-set clustering used to precluster the most informative feature vectors.
- The MII criteria takes into account high-order feature interactions, overcoming the problem of overestimated redundancy. As a result, the feature components associated with the greatest amount of joint information can be preserved.

Future work

- Represent feature-subsets using hypergraphs.
- Alternative information measures.
- Better search.

Reference

[1] S.Buchala, N.Davey, T.M.Gale, and R.J.Frank. Principal component analysis of gender, ethnicity, age, and identity of face images. In Proc, IEEE Int'l Conf.Multimodel Interfaces, 2005.

[2] R.Battiti, Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on neural networks 5(4), 537-550 (2002)

[3] H.Peng, F.Long, C.Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance , and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence pp. 1226-1238 (2005)

[4] H.Yang, J.Moody, Feature selection based on joint mutual information. In proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis, pp. 22-25 (1999)

[5] M.Pavan, M.Pelillo, A new graph-theoretic approach to clustering and segmentation. In IEEE computer society conference on computer vision and pattern recognition. Vol.1. IEEE (2003)

Thanks ! And Questions ?