

Towards Structural Feature Selection

Fabrizio Costa and Björn Bringmann

Katholieke Universiteit Leuven, Belgium
{Fabrizio.Costa,Bjoern.Bringmann}@cs.kuleuven.be

Abstract. This work explores the idea of selecting features based on their internal structure. We introduce a redundancy notion for features based on their structural similarity and cast the problem of selecting the optimally non-redundant set of b features as a variant of a weighted maximal b -clique problem. We propose a greedy algorithm suitable for large problem sizes. The original features are patterns correlated with the target class extracted using a pattern miner. Results on molecular datasets show that this method allows us to decrease the number of features to 10% of the original size without a significant decrease in the classification performance for a range of diverse classifiers.

1 Structural Feature Induction

Graph miners are one of the most prominent of structured data mining systems. Their application has been to a large extent targeted towards problems in the molecular domain such as structure activity relationship prediction [5]. Typically classification approaches involving pattern mining techniques use a set of patterns extracted by a mining task to transform each instance into a fixed-length binary vector. Each bit in this vector corresponds to the occurrence of a pattern in the instance. Due to the complexity of structured data such as molecules, expressive pattern languages to represent sequences, trees or graphs are employed. This increased expressivity comes at the price of obtaining very similar, hence, we argue, redundant, features posing a selection problem that needs to be addressed. Empirical evidence from feature selection literature shows in fact, that redundant features affect the speed and accuracy of classification algorithms and should therefore be eliminated together with the irrelevant ones.

While there is no natural way to compare standard scalar features without resorting to their occurrences in some instances of a given problem, structured features exhibit a rich internal representation that can be independently exploited to assess their relative *structural similarity* by means of appropriate kernel functions. Here we assume that a feature is good if it is relevant to the class concept but is not redundant with respect to any of the other features. While the relevance to the class is achieved by mining correlated patterns as in [2] we define redundancy as structural similarity. The two contributions of this work are the definition of feature redundancy as a kernel function specialized for the case of small molecules and a fast algorithm to select a fixed size subset of features that minimizes the cumulative pairwise similarity.

Algorithm 1 Structural redundancy reduction

- 1: **Input:** Set of graphs $S = \{x_1, \dots, x_k\}$, Graph kernel $K(\cdot, \cdot)$, size of the sub-set b , number of random swaps m
 - 2: **Return:** List of b graphs $O = (x_i, \dots, x_j)$ maximally dissimilar
 - 3: Initialize O with the maximally dissimilar pair (\hat{x}_1, \hat{x}_2)
 - 4: **for** b times **do**
 - 5: $\arg \min_{x_i \in S} \sum_{x_j \in O} K(x_i, x_j)$
 - 6: add x_i to O and remove it from S
 - 7: **for** m times **do**
 - 8: derive O' and S' by swapping $x_i \in O \leftrightarrow x_j \in S$
 - 9: **if** $\sum_{x_j \in O'} K(x_i, x_j) < \sum_{x_j \in O} K(x_i, x_j)$ **then**
 - 10: update $O \leftarrow O'$ and $S \leftarrow S'$
-

2 Redundancy as Similarity

The first contribution is an interpretation of redundancy as a kernel between features. We use the weighted decomposition kernel (WDK), a specialization of a decomposition kernel [4], originally introduced in [9]. Decomposition kernels are a class of kernels built as composition of simpler kernels over fragments or parts of a complex data type such as a sequence or a graph. A WDK specializes the decomposition kernel by introducing weights for the fragments computed by a probability kernel [7] over the *contexts* of the fragments. We adapt the WDK to deal with small molecules by considering the atoms as vertices and the chemical bonds as edges. Each atom v is characterized by a small region of topological *nearby* elements – the context $\mathcal{C}_x^l(v)$ – defined as the sub-graph composed of the vertices within l hops from vertex v . We compute the similarity between two molecules in terms of the similarity of the set of their vertices \mathbb{V}_x weighted by the atom types present in their respective neighborhoods as: $K(x, x') = \sum_{v \in \mathbb{V}_x, v' \in \mathbb{V}_{x'}} \delta(v, v') \cdot k(\mathcal{C}_x^l(v), \mathcal{C}_{x'}^l(v'))$. The exact matching kernel $\delta(v, v')$ is 1 if vertices v and v' have the same label (atom type in our case). The context kernel $k(\cdot, \cdot)$ is the scalar product of the histograms built collecting atom type and edge type information (a compound of bond type *and* both incident vertices types) over the nodes that are part of the context.

3 Redundancy Reduction

Given a similarity Gram matrix associated to a set of features and equating similarity with redundancy we can reformulate the problem of redundancy reduction as a weighted maximal b -clique problem (WCP $_b$) [8]: given an undirected graph with weights on edges¹ the task in WCP $_b$ is to find a clique with at most b nodes such that the sum of all the weights in the subgraph is maximal. In our case the edge weight function is obtained from the WDK kernel.

¹ In the general WCP $_b$ formulation there exist a weight function for the vertices too, however in our case we assume all vertices to have equal unit weight.

The weighted maximal b -clique problem is NP-hard and integer-programming solutions even for relaxed approximations are suitable only for small size problems (tens of vertices) [6]. In bio-informatics applications the number of variables typically ranges in the interval of thousands to tens of thousands. Therefore we provide an algorithm applicable to such large-scale problems when no exact solution is required as our second contribution.

We propose a simple strategy based on a greedy approach and an iterative random refinement (see Algorithm 1). Starting from the most dissimilar pair we increases a working set of least structural-redundant candidates in an incremental way. The algorithm uses a set S initialized with all patterns and a set O initialized with the (selected) most dissimilar patterns. A total of b elements are iteratively removed from S and inserted into O . In each step the element from S with the maximally cumulative dissimilarity (i.e.: the lowest sum of the projections defined by the kernel) in O is selected. Finally O is refined by randomly swapping elements in O with elements in S if by doing so the cumulative dissimilarity of O is improved. Good values for the selected number of patterns b can be determined by means of cross-validation techniques once the selected sub-set of features is used in a classification task.

The complexity of the algorithm, dominated by the computation of the cumulative dissimilarity, is quadratic in the size of the selected subset allowing its usage for many real world bio-informatics applications.

4 Preliminary Experimental Results and Conclusions

We evaluated the approach on three datasets with approximately 700, 1000, and 3000 chemical compounds respectively: **PTC** (carcinogenicity properties on rodents); **AID651** (a HIV related bio-assay); and the **NCI** Cancer Dataset (screening results of ability to suppress or inhibit growth of various tumor cell lines, modified and made available from the ChemDB project). We explored the effect of using a subset of the 1K most correlated ($\chi^2 \geq 3.84$) patterns² on a range of well known binary classifiers: naïve bayes, k-nearest neighbor, RIPPER [3], and an SVM with a Tanimoto kernel [1]. We compare to a measure of similarity obtained by representing the features with a one-hot-encoding of the instances were they appear. We compute the similarity matrix in terms of the dot product in the associated vector space and seamlessly employ Algorithm 1 for the redundancy reduction phase. Given that the two similarities are uncorrelated (as evaluated by the Spearman Rank Correlation test), we also confronted the similarity measure obtained as a linear combination of the structure-based and the instance-coverage-based Gram matrices. We measured the AUC for the ROC over 5-fold stratified cross-validated experiments. Results indicate that classifiers using as few as 10% of the originally mined patterns do not significantly exhibit a degradation in performance. All classifiers, apart from SVM, show indeed an increased average performance (although not significantly) compared to the case where all top 1000 correlated patterns are used as features. In several cases we

² Based on experiences from [2] we used free *sequences* as pattern language with a partial ordering based on subgraph-isomorphism.

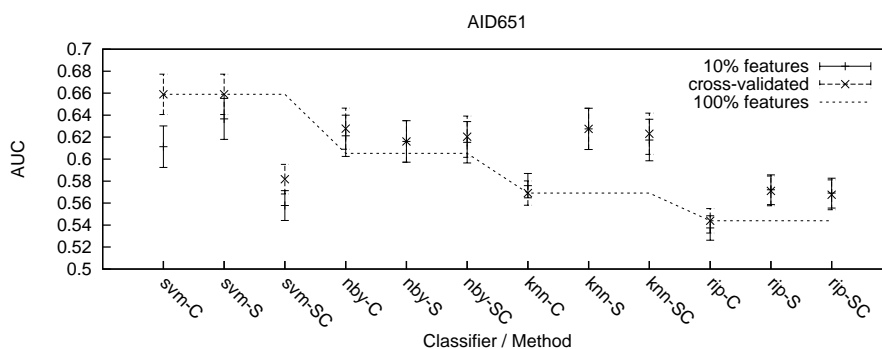


Fig. 1. AUC for S)tructural, C)overage and combined SC) similarity based reductions

noticed that by using structural similarity we could improve over the coverage similarity (significantly better for RIPPER and KNN, better but not significantly for SVM, and not significantly worse for naïve bayes). Surprisingly, the combination of the two similarity measures does not improve over the single measures as their uncorrelatedness would suggest. Finally we re-confirm how the selection of informative features can make weaker classifiers as naïve bayes and k-nearest neighbor reach the performance level of kernel based SVMs. These experiments are still in a preliminary phase and a more thorough investigation of the influence of specific kernels in the feature selection process is due.

Acknowledgment

The authors would like to thank Marco Signoretto for suggestions and fruitful discussions.

References

1. Fligner M. A, Verducci J. S, and Blower P. E. A modification of the jaccard-tanimoto similarity index for diverse selection of chemical compounds using binary strings. *Technometrics*, 44:110–119, May 2002.
2. Björn Bringmann, Albrecht Zimmermann, Luc De Raedt, and Siegfried Nijssen. Don't be afraid of simpler patterns. In *PKDD*, pages 55–66, 2006.
3. William W. Cohen. Fast effective rule induction. In Armand Prieditis and Stuart J. Russell, editors, *ICML*, pages 115–123, July 1995.
4. D. Haussler. Convolution kernels on discrete structures. Technical Report UCS-CRL-99-10, UC Santa Cruz, 1999.
5. C. Helma, T. Cramer, S. Kramer, and L. De Raedt. Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. *Journal of Chemical Information and Computer Systems*, 44(4):1402–1411, 2004.
6. M. Hunting, U. Faigle, and W. Kern. A lagrangian relaxation approach to the edge-weighted clique problem. In *European Journal of Operational Research*, volume 131, pages 119–131, 2001.
7. T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *J. Mach. Learn. Res.*, 5:819–844, 2004.
8. E. M. Macambira and C. C. de Souza. The edge-weighted clique problem: Valid inequalities, facets and polyhedral computations. In *European Journal of Operational Research*, volume 123, pages 346–371, 2000.
9. S. Menchetti, F. Costa, and P. Frasconi. Weighted decomposition kernels. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.