A Feature Selection Method based on Tree Decomposition of Correlation Graph

Abdelkader Ouali, Nyoman Juniarta, Bernard Maigret, Amedeo Napoli

A Feature Selection Method based on Tree Decomposition of Correlation Graph

Abdelkader Ouali[1], Nyoman Juniarta[1], Bernard Maigret[1], and Amédéo Napoli[1]

Université de Lorraine, CNRS, Inria, LORIA F-54000 Nancy, France.
firstname.lastname@loria.fr

Molecular descriptors are tightly connected to the concept of molecular structure and play an important role in the scientific sphere due to their connection to a complex network of knowledge [7]. The availability of large numbers of theoretical descriptors, currently by thousands, provides diverse sources to better understand relationships between molecular structure and experimental evidence, but adds more complexity by introducing new features. To confront the high dimensionality of data, a correct selection of the features allows to reduce the size of the data and the computational time while preserving the predictive power of the classifier. Feature selection methods, wherein subsets of features available from the original data are selected, are now commonly used in data mining and represent an active field of research [3]. Some of these methods rely on correlation graph to reveal structural properties hidden in the data. These graph-based approaches belong to the class of filter-based methods. We introduce a method which uses a tree decomposition [4] of feature-feature correlation graph to select representatives in order to reduce redundancy. This method is exploited to target molecular descriptors involved in identifying molecules that can be characterized as antibacterials.

Prior Work. Different graph-based approaches have been proposed for feature clustering where the main objective is to avoid the selection of redundant features. The authors in [6] use graph-theoretic clustering methods to separate features into clusters, representatives are selected next using features which are strongly related to target classes. The authors in [8] use a hyper-graph clustering to extract maximally coherent feature groups from the dataset which includes third or higher order dependencies. The authors in [1] proposed a method which relies on community detection techniques to cluster graphs which describe the strongest correlations among features. Compared to existing graph-based feature selection methods, our method exploits tree decomposition techniques to reveal groups of highly correlated features.

FSTDCG[1] method. A dataset $\mathcal{D}$ is described as set of examples $\mathcal{M}$ and a set of features $\mathcal{F}$. Each example $m \in \mathcal{M}$ is defined w.r.t. all the features in $\mathcal{F}$. Let $Dom(f)$ be the set of values taken by a feature $f \in \mathcal{F}$ over all the examples in $\mathcal{M}$. We assume that Dom is a numerical set for all features in $\mathcal{F}$. One specific feature can act as the *class* of the examples. We assume that the set of values of a feature class $Dom(class)$ is defined in $\mathbb{N}$. A feature selection to $\mathcal{F}$, resulting in $\mathcal{F}_S$, is the set of more discriminating features for the classes in the dataset. Our method proceeds in three steps. First, a correlation graph is computed between each two features $(x, y) \in \mathcal{F}$ using the Pearson Product

[1] Feature Selection Tree Decomposition Correlation Graph.

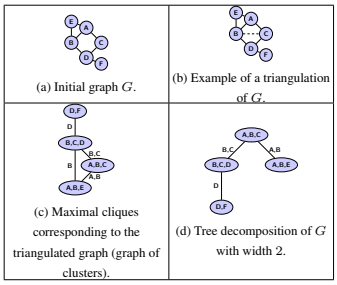

Fig. 1: Steps for computing a tree decomposition of a graph G

Algorithm 1: Pseudo code of FSTDCG.

Input: A set of features $\mathcal{F}$, and Correlation Graph G.
Output: Selected subset of features $\mathcal{F}_S$.

1 $\mathcal{F}_S \leftarrow$ `notCorrelatedFeatures` $(\mathcal{F}, G)$;
2 $(C_T, T) \leftarrow$ `treeDecomposition` (G);
3 **for each** *cluster* $C \in C_T$ **do**
4 $f_{max} \leftarrow 0, max \leftarrow 0$;
5 **for each** *feature* $f \in C$ **do**
6 **if** `Occur` $(f, C_T) > max$ **then**
7 $max \leftarrow$ `Occur` (f, C_T);
8 $f_{max} \leftarrow f$;
9 $\mathcal{F}_S \leftarrow \mathcal{F}_S \cup \{f_{max}\}$;
10 **return** $\mathcal{F}_S$;

Moment Correlation [5]. An edge is created for each (x, y) having a correlation value higher than empirical threshold. Second, a tree decomposition of the correlation graph is computed based on the min-fill heuristic [2]. Fig. 1 depicts the different steps (a-d) to compute the tree decomposition. The clusters C_T produced in step (c) gather highly correlated features since clusters correspond to maximal cliques which cannot be extended by including one more adjacent node. Third, representatives are selected based on their maximum occurrence in other clusters.

Algorithm 1 depicts the pseudo code of FSTDCG method. The uncorrelated features not found in the correlation graph are kept in the set $\mathcal{F}_S$, see line 1. FSTDCG selects the feature with a maximum occurrence over all the clusters of the tree decomposition. This strategy allows to benefit from the topology of the tree, and to reduce redundancy among features appearing in different clusters, see lines 3-9.

Experiments. Experiments were carried out on 18 different datasets available from the UCI repository. An edges is created in the correlation graph if the ρ-value of the correlation is greater than 0.05. We used Random Forest classifier with 10-fold cross validation to evaluate the accuracy. The method provides a reasonable number of features, on average 20.32% of the original datasets, while decreasing on average 4.36% of the accuracy of the classifier. We are currently investigating the use of FSTDCG to address more specific molecular datasets. Further work will consist of using evolve graphs to handle changes that occur on the dataset when new data are added or removed.

References

1. Horvath, S.: Correlation and Gene Co-Expression Networks (2011)
2. Kjaerulff, U., Datasystemer A/s, J.: Triangulation of graphs - algorithms giving small total state space (2002)
3. Li, J., Liu, H.: Challenges of feature selection for big data analytics (2017)
4. Robertson, N., Seymour, P.D.: Graph Minors. II. Algorithmic Aspects of Tree-Width (1986)
5. Rodgers, J.L., Nicewander, W.A.: Thirteen ways to look at the correlation coefficient (1988)
6. Song, Q., Ni, J., Wang, G.: A fast clustering-based feature subset selection algorithm for high-dimensional data (2013)
7. Todeschini, R., Consonni, V.: Molecular Descriptors for Chemoinformatics (2009)
8. Zhang, Z., Hancock, E.R.: A graph-based approach to feature selection (2011)