**Abstract**

配合相关中文科普食用更佳

feature selection techniques in bioinformatics

taxonomy + usage

**1 Introduction**

feature selection (FS) techniques：a real prerequisite for model building

FS techniques do not alter the original representation of the variables, but merely select a subset of them → they preserve the original semantics of the variables, hence, offering the advantage of interpretability by a domain expert

Application：supervised (classification; the class labels are known beforehand) and unsupervised (clustering; more complex) learning

**2 Feature Selection Techniques**

与pattern recognition techniques联用

Manifold objects：

(a) to avoid overfitting and improve model performance, i.e. prediction performance in the case of supervised classification and better cluster detection in the case of clustering

(b) to provide faster and more cost-effective models and

(c) to gain a deeper insight into the underlying processes that generated the data

与其优化整个特征集的最优模型参数，不如寻找最优特征子集的最优模型参数Instead of just optimizing the parameters of the model for the full feature subset, we now need to find the optimal model parameters for the optimal feature subset, as there is no guarantee that the optimal parameters for the full feature set are equally optimal for the optimal feature subset

与其他技术的不同：incorporate this(以finding the optimal subset of relevant features为model hypothesis space的拓展) search in the added space of feature subsets in the model selection

three categories (depending on how they combine the feature selection search with the construction of the classification model): filter methods, wrapper methods and embedded methods.

*Filter techniques*

looking only at the intrinsic properties of the data, low-scoring features are removed, features is presented as input to the classification algorithm → easily scale to very high-dimensional datasets, simple and fast, independent of the classification algorithm, be performed only once; ignore the interaction with the classifier, univariate → multivariate filter techniques

*wrapper methods*

embed the model hypothesis search within the feature subset search

The evaluation of a specific subset of features is obtained by training and testing a specific classification model, rendering this approach tailored to a specific classification algorithm

as the space of feature subsets grows exponentially with the number of features, heuristic search methods are used to guide the search for an optimal subset

优点：interaction between feature subset search and model selection, and the ability to take into account feature dependencies

缺点：higher risk of overfitting, computationally intensive

*embedded techniques*

the search for an optimal subset of features is built into the classifier construction

优点：interaction with the classification model, less computationally intensive

**3 Applications in Bioinformatics**

**3.1 Feature selection for sequence analysis**

two types of problems: content (broad characteristics) and signal (identification of important motifs) analysis

除了上述基础特征，还有higher order combinations (e.g. k-mer patterns)

**3.1.1 Content analysis**

预测蛋白质序列与功能（例：recognition of promoter regions, prediction of microRNA targets）

*结构*：

MM: many features can be extracted from a sequence, and most dependencies occur between adjacent positions

IMM: the high amount of possible features, and the often-limited amount of samples → future：deal with non-adjacent feature dependencies (ICM: crosses a Bayesian decision tree with a filter method(χ2))

MBF

*功能*：

SVM: subsequently remove features with low weights

**3.1.2 Signal analysis**

the recognition of short, more or less conserved signals, 代表结合位点

Feature selection can then be used to search for the motifs that maximize the fit to the regression model

TNoM: to score genes for relevance to tissue classification

翻译起始位点（TIS）、剪切位点等结构元素（structural elements）预测

EDA：获取更多相关特征

**3.2 Feature selection for microarray analysis**

微阵列（microarray）分析：数据量大、样本数少 → 降维技术尤为重要

**3.2.1 The univariate filter paradigm: simple yet efficient**（缺点：未考虑基因间相互作用）

微阵列分析只用单变量方法, two sample t-test and ANOVA are among the most widely used techniques in microarray studies

new or adapted univariate feature ranking techniques has since then been developed: parametric and model-free methods: Gaussian assumptions, regression modelling approaches, Gamma distribution models; Wilcoxon rank-sum test, BSS/WSS, rank products method → 增强了对异常值的鲁棒性，特别关注控制错误发现率(FDR)的贡献

**3.2.2 Towards more advanced models: the multivariate paradigm for filter, wrapper and** **embedded techniques**

相关方法：CFS, Markov blanket filter method, MRMR, USC, sequential search techniques, approximation of the area under the ROC curve (近期倡导), the optimization of the LASSO model (近期倡导)

推荐：to pre-reduce the search space using a univariate filter method, and only then apply wrapper or embedded methods, hence fitting the computation time to the available resources.

**3.3 Mass spectra analysis质谱分析**

The data analysis step is severely constrained by both high-dimensional input spaces and their inherent sparseness, just as it is the case with gene expression datasets.

与microarray一样，univariate filter techniques最为常见

**4 Dealing with small sample domain**

Two initiatives: 1) the use of adequate evaluation criteria

2) the use of stable and robust feature selection models.

**4.1 Adequate evaluation criteria**

为保证准确性，在训练过程的每个阶段都引入一个外部特征选择过程

novel predictive accuracy estimation methods: bolstered error estimation

**4.2 Ensemble feature selection approaches**

methods based on a collection of decision trees (e.g. random forests)

集成方法与技术

**5 Feature Selection in Upcoming Domains**

**5.1 Single nucleotide polymorphism analysis (SNPs)**

基因-疾病关联，完整基因型预测

selecting a subset of SNPs that is sufficiently informative but still small enough to reduce the genotyping (基因分型：基因分型是通过使用生物学试验检查个体的DNA序列的过程) overhead is an important step towards disease-gene association

Genetic algorithms have been applied to the search of the best subset of SNPs, evaluating them with a multivariate filter (CFS), and also in a wrapper manner, with a decision tree as supervised classification paradigm

**5.2 Text and literature mining**

BOW: 文本中的每个词表示一个变量，其值由文本中特定单词的频率组成. 虽然特征选择技术在文本分类领域的应用很普遍，但在生物医学领域的应用仍处于起步阶段

e.g. Kullback–Leibler divergence as a univariate filter method to find discriminating words in a medical annotation task; symmetrical uncertainty (an entropy-based filter method) for identifying relevant features for protein interaction discovery; the use of feature selection for a document classification task

**6 FS Software Packages**

Four sections: 1) general purpose FS techniques

2) techniques tailored to the domain of microarray analysis

3) techniques specific to the domain of mass spectra analysis

4) techniques to handle SNP selection

**7 Conclusion and Future Perspectives**

Two main issues emerge as common problems in the bioinformatics domain: the large input dimensionality, and the small sample sizes. → FS techniques

univariate filter FS 有一定局限性，最好考虑multivariate selection algorithms1

集成2，提高鲁棒性（在异常和危险情况下系统生存的能力）

向生物信息学3方向发展：SNPs, text and literature mining, combination of heterogeneous data sources