第 33 卷 第 8 期 计 算 机 工 程 2007 年 4 月
Vol.33 No.8 Computer Engineering April 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)08—0182—03 文献标识码：A 中图分类号：TP181

FSL-SP 的研究

施 佳，夏骄雄，张 武
(上海大学计算机工程与科学学院，上海 200072)

摘 要：在机器学习领域，特征选择对于提高学习机器的性能和效率具有重要意义，但是当前特征选择算法普遍存在着具体实现独立性强、可扩展性差的问题，使得对多种算法性能的统一对比评估实施困难，算法的替换和扩展比较复杂。以面向对象的设计理念为指导，基于设计模式中的策略模式，提出特征选择算法工具库 FSL 的设计构想，通过将一些常用的特征选择算法按照策略模式进行包装，以便机器学习算法用户的使用，同时确保其较强的可扩展性。

关键词：机器学习；特征选择；策略模式；FSL-SP

Research on FSL-SP

SHI Jia, SHARDROM Johnson, ZHANG Wu
(School of Computer Engineering and Science, Shanghai University, Shanghai 200072)

【Abstract】 Feature selection is very important in improving the performance of learning systems. Various feature selection algorithms greatly facilitate the research of the scientists from different disciplines, there is a common problem that those algorithms are implemented by different researchers. So it is hard for the users to integrate or compare those independent implementations of different programming styles and incompatible designs. The feature selection library on strategy pattern(FSL-SP) is conceived to solve the above problem. The FSL-SP encapsulates many popular feature selection algorithms under unified interfaces, while different strategies of one algorithm could be exchanged conveniently. This library will bring help to those machine-learning algorithms users. The FSL-SP itself has good extensibilities, and new algorithm can be added into the library easily.

【Key words】 Machine learning; Feature selection; Strategy pattern; Feature selection library on strategy pattern (FSL-SP)

机器学习作为人工智能研究的重要分支，其各类算法在众多领域得到广泛应用[1]。随着学习样本数据量的不断增大，海量非规则信息对机器学习的训练和预测所需的计算能力和存贮容量提出更高的要求。特征选择技术正是顺应这一需求，以提高机器学习的预测性能为目标，提供理解和探求样本数据产生过程的有效手段[2]。

虽然特征选择算法对提高学习机器的性能具有针对性[3]，但其具体实现的独立性强、可扩展性弱的特点，导致难以实施统一的性能对比评估，且替换和扩展其中相关算法较复杂。为此，本文提出基于策略模式的特征选择算法工具库(Feature Selection Library on Strategy Pattern，FSL-SP)的设计构想，利用特征选择过程所具备的良好模块化特性，根据面向对象领域的设计理念，将一些常用的特征选择方法抽象为统一的包装接口，适应较强可扩展性的需求，有利于算法部分的修改和调整。

1 设计模式

面向对象的设计模式[4]应用正成为软件开发的重要组成部分。FSL-SP采用设计模式中的策略模式进行各种特征选择算法的策略抽象组织，以便实现策略互换，具备良好的可扩展性。基于“一个策略对应一种方法或算法”的原则，策略模式用于对一系列同类算法进行包装，并将算法的使用与实现分离，将各种算法委派给不同的对象管理[5]。例如，将一系列算法实现包装到相应的策略类中，策略类之间可以进行互换以适应不同用户的需求，并引入一个抽象类作为这些策略类的基类。

如图 1 所示的策略模式实现图中，ConcreteStrategy 代表具体的策略类，多个策略类同时并存，且易于互换，均继承自抽象类 Strategy，并以合成关系包含在上下文 Context 中。

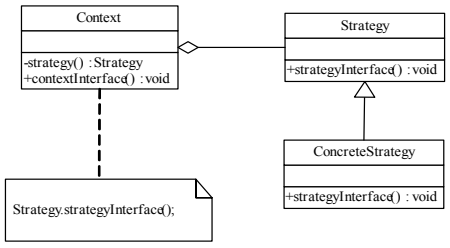

图 1 策略模式的基本结构

面向对象的软件系统中类与类之间的区别若仅限于它们的行为，则可以将这些行为(对应某种算法)包装成相应的具体类，使之统一继承于一个抽象基类。根据面向对象的编程原则和多态性原理，用户通过基类来动态选择任一子类的具体算法，避免接触那些只与算法有关的数据，可以大大降低客户代码与算法实现代码之间的耦合度。如果一个对象有许多行为，使用策略模式可以在实现中避免使用多重条件选择语句，体现面向对象的设计理念，有利于构造一个设计优良、可扩展性强的软件系统。

基金项目：上海市高等学校科学技术青年基金资助项目(01QN59)；上海市高等学校科学技术发展基金资助项目(04AB29)
作者简介：施 佳(1981－)，男，硕士生，主研方向：机器学习，数据挖掘；夏骄雄，博士、讲师；张 武，博士、教授、博导
收稿日期：2006-04-27 **E-mail：**shuwizard@gmail.com

2 FSL-SP 的设计

FSL-SP 的设计构想源于特征选择算法所具有的良好模块化特性，把抽象得到的众多子模块通过多种方法加以实现，即选取各种外部效果近似而内部形式不同的算法来实现。

2.1 特征选择方法的分类

特征选择过程大多从样本记录的属性全集中筛选出实际用于预测的特征子集，这种子集选择法大致分为Filter、Wrapper和Embeded方法[3]。Filter方法独立于预测用的学习机器，只对数据记录属性之间的信息进行特征选择；Wrapper方法按照某种搜索办法获得特征属性集作为学习机器的输入，并根据预测的准确率来评估特征，确定取舍并完成迭代[6]；Embeded方法则在每次迭代时都使用特征集进行训练，针对某种学习机器优化其预测性能，还不需要在每次迭代时划分训练样本集与测试样本集。

FSL-SP基于Guyon提出的方法对特征选择方法进行分类，定义相应的策略(方法)[3]，并遵循面向对象的理念设计类的继承架构，从而有效成为实现FSL-SP设计的根本出发点，即统一包装这些策略并方便替换使用。一方面，确定搜索所有可能特征子集的办法，在样本属性集较小的情况下选择穷举搜索的办法，但随着属性集的变大、搜索问题的计算复杂度增加，采用其他的搜索策略。另一方面，选择一种学习机器(预测器)，对每次预测的结果进行合理评价，决定特征的取舍，并控制搜索的实施。

2.2 特征选择算法的策略抽象

FSL-SP 将整个特征选择过程分解成一些抽象的步骤，每个步骤的实现依赖于包装各种策略的核心功能模块，模块内的策略可以进行互换。基于 FSL-SP 设计的特征选择方法可以在不同的步骤，根据具体所需获取相应模块，定制具体的算法流程。

(1)初始阶段，基于特征选择方法的类别选取划分策略将样本数据分为训练数据和测试数据，并按照选定的特征搜索策略将样本记录的所有或部分属性加入候选特征集。

(2)训练阶段，基于特征选择方法的类别训练选用的预测机器。

(3)特征评价阶段，选择一种特征评价方法，根据评价结果对特征属性进行处理，同时根据第(1)步选定的特征搜索策略决定特征的取舍。

(4)结束阶段，根据用户指定的要求来决定整个特征选择过程，最后返回特征集作为输出结果。

这种细粒度的划分有利于理解特征选择算法的机制，清晰特征选择算法的设计思路。组件化定制特征选择算法的选择过程，使 FSL-SP 用户大大加快和简化开发与使用过程。表 1 示意了特征选择算法流程各功能模块的策略。

表 1 特征选择算法流程各功能模块及策略

抽象步骤	核心模块	可供选择的各类策略
初始阶段	样本数据划分策略模块	二重随机划分，M-重划分等
	搜索策略模块	最佳匹配，模拟退火，遗传算法，贪心算法等
训练阶段	预测器选择策略模块	决策树，朴素贝叶斯，支持向量机等
	预测器验证策略模块	SINGLE VALIDATION，LEAVE-ONE-OUT 等
特征评价阶段	FILTER 特征选择策略模块	PEARSON 方法，FISHER 方法等
	WRAPPER/EMBEDED 特征选择策略模块	MOODY 基于预测风险的评价方法，OBD 等
结束阶段	终止判断策略模块	用户自定义

以基于预测风险的特征选择算法SVM-SBS[7]为例。根据FSL-SP中特征选择方法的分类，属于Embeded方法的SVM-SBS通过直接训练多类SVM确定输入数据的特征集合，并根据Moody提出的基于预测风险(prediction risk based)特征评价函数[8]进行评价：

$$S_i = ERR(\overline{x}^i) - ERR$$

其中，ERR 为多类 SVM 的训练误差；

$$ERR(\overline{x}^i) = \frac{1}{N}\sum_{j=1}^{N}(\tilde{y}(x_j^1,\ldots,\overline{x}^i,\ldots,x_j^M) \neq y_j)$$

其中，M、N 为特征数和训练样本数。

根据 FSL-SP 定制的 SVM-SBS 算法将某个特征属性 x^i 用其在整个样本集的平均值 $\overline{x}^i$ 代替，通过计算替代前后的训练误差变化大小来决定特征的取舍。其基本流程为：采取后向搜索策略将样本记录所有属性加入候选特征集 U，将剔除特征集 R 和测试错误列表 E 置空，然后按某种划分策略将样本数据划分为训练数据和测试数据；选择并通过训练得到一个多类 SVM 预测分类器，测试分类器并计算分类误差，保存在列表 E 中；根据 Moody 方法特征评价函数确定评价结果值最小的特征属性 h，放入剔除特征集 R 中；评估没有 h 的候选特征集，若特征集的大小不满足用户指定的需求则返回第(1)步，否则返回候选特征集作为结果。

上述流程的各个步骤可以抽象为模块，每一个策略模块可以采用不同的方法实现。在 FSL-SP 的设计实现中，这些不同方法都将抽象成具体的策略类，符合继承和合成关系，并可以按照具体需求实施替换。

2.3 FSL-SP 架构设计

FSL-SP 以策略模式的设计思想为指导，采用面向对象的方法，根据特征选择方法的分类和特点，设计工具库的架构。每一个特征选择方法，包括整个特征选择过程中使用的策略，都包装在一个类中。特征选择过程中选用的大量算法根据功能区别分类，并包装在单个策略类中。核心模块对应一个策略基类，功能相近的各种具体策略类实现相关的算法，并继承自同一个策略基类。策略类的对象以合成的关系放在一个特征选择方法类中。设计完成一个特征选择方法，用户可以将选定的各种所需策略传入类的构造函数中，从而得到一个定制的特征选择方法类。

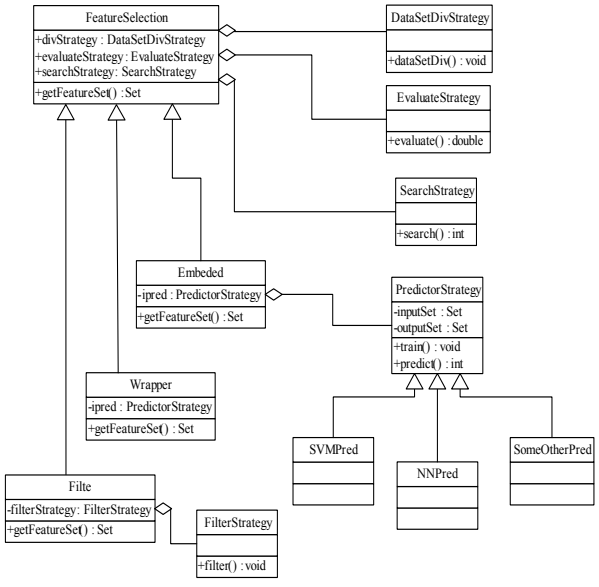

图 2 FSL-SP 的架构设计

图 2 显示库中主要抽象类的继承和合成关系。FeatureSelection 为 FSL-SP 的核心基类，根据合成关系封装策略基类的对象，包括样本数据集划分策略(DataSetDivStrategy)、特征空间搜索策略(SearchStrategy)、特征评价策略(EvaluateStrategy)。GetFeatureSet 作为特征选择过程统一抽象方法，根据初始输入返回用户所需的特征集合。Filter、Wrapper 和 Embeded 都作为 FeatureSelection 的子类，对应 3 类特征选择方法，并各自封装策略对象(FilterStrategy)和学习机器策略(PredictorStrategy)。PredictorStrategy 可以派生出各种不同类型的学习机器，并在保持接口一致的情况下发展各自的继承结构。特征选择过程各个步骤涉及的算法都在 FSL-SP 的实现中归类至相应的策略类中，用户可以在构造特征选择工具时方便地选用。

3 基于 FSL-SP 的特征选择方法实现

基于 FSL-SP 构造的 SVM-SBS 特征选择算法基本结构如下：

```
class SVMSBS : public Embeded
{
  public:
        SVMSBS(DataSetDivStrategy*,        // 数据集划分策略
PredictorStrategy*,                 // 分类器选择策略
SearchStrategy*,                    // 特征属性搜索策略
EvaluationStrategy*,        // 特征评价策略
QuitStrategy*,                              // 迭代中止判断策略
Set);                          // 输入属性全集(候选特征集)
Set getFeatureSet();  // 统一特征获取方法
…
};
Set SVMSBS::getFeatureSet()
{
/*dsDivStrategy.dataSetDiv(…);*/              // 样本集划分方法
  while (quitStrategy->satisfied(…)){    // 自定义终止条件
        ipred->train(…);                          // 确定学习机器
        evaluateStrategy->evaluate(…); // 确定特征评价方法
update(featureSet);                         // 确定特征属性的取舍
searchStrategy->search(…);        // 确定取舍特征搜索策略
  }
  return featureSet;
}
```

在 FSL-SP 中，策略类的实现都隶属于同一模块，用户只需通过向构造函数传递不同的策略对象就能够定制特征选择过程的各个核心模块，以及进行两种策略或者多种特征评价方法在特定数据集上的表现比较。由于统一的接口标准，FSL-SP 具有良好的可扩展性，各种子模块的新算法只需要遵循接口标准包装算法即可实现扩展。

4 结论

基于策略模式的特征选择算法工具库 FSL-SP 可以明显提高特征选择过程中各种算法的重用率，并且具有较强的可扩展性。它所提供的细粒度可定制特征选择方法构造解决方案，可以直接定制需要的特征选择方法，并通过比较不同具体策略的实际运行性能，决定适合特定训练样本集的选择，大大节省修改代码的时间花费。当然，FSL-SP 要求用户必须熟悉所有策略类，并自行决定具体策略类的使用，这限定了用户的适用范围。而且，各种不同的特征选择方法虽然具有明显的模块化特征，但是有关精确划分模块的办法还需进一步研究。

参考文献

1 Tom M M. Machine Learning[M]. New York: McGraw-Hill, 1997.

2 Vladimir N V. The Nature of Statistical Learning Theory[M]. Berlin: Springer-Verlag, 1995.

3 Isabelle G, Andre E. An Introduction to Variable and Feature Selection[J]. Journal of Machine Learning Research, 2003, 3(1): 1157-1182.

4 Erich G. Design Patterns: Elements of Reusable Object-oriented Software[M]. Boston: Addison-Wesley Professional, 1995.

5 阎　宏. Java 与模式[M]. 北京: 电子工业出版社, 2002.

6 Ron K, George H J. Wrappers for Feature Subset Selection[J]. Artificial Intelligence, 1997, 97(1/2): 273-324.

7 Li Guozheng, Jie Yang. Feature Selection for Multi-class Problems Using Support Vector Machines[EB/OL]. http://cs.shu.edu.cn/gzli/publication.htm.

8 John E M, Joachim U. Principled Architecture Selection for Neural Networks: Application to Corporate Bond Rating Prediction[C]//Proc. of Advances in Neural Information Processing Systems. 1992.

(上接第 179 页)

参考文献

1 Sofien T, Najoua E B A, Hamid A. Generalized Hough Transform for Arabic Optical Character Recognition[C]//Proceedings of the 7th International Conference on Document Analysis and Recognition. 2003.

2 Li Zenian, Osmar R Z, Zinovi T. Illumination Invariance and Object Model in Content-based Image and Video Retrieval[J]. Journal of Visual Communication and Image Representation, 1999, 10(3): 219–244.

3 Rong C L, Wen G T. Perspective-transformation-invariant Generalized Hough Transform for Perspective Planar Shape Detection and Matching[J]. Pattern Recognition, 1997, 30(3): 383-396.

4 Cózar J R, Guil N, Zapata E L. Detection of Arbitrary Planar Shapes with 3D Pose[J]. Image and Vision Computing, 2001, 19(14): 1057-1070.

5 Marcus A M. Geometry-based Automatic Object Localization and 3-D Pose Detection[C]//Proc. of IEEE Southwest Symposium on Image Analysis and Interpretation. 2002.

6 Witold Ż, Brian F, Johnathan B. Irregular Colour Pattern Recognition Using the Hough Transform[C]//Proc. of the 3rd IMA Conference on Imaging and Digital Image Processing: Mathematical Methods, Algorithms and Applications. 2000.

7 Pui K S, Wan C S. A New Generalized Hough Transform for the Detection of Irregular Objects[J]. Journal of Visual Communication and Image Representation, 1995, 6(3): 256-264.

8 Chau Chunpong, Siu Wanchi. Adaptive Dual-point Hough Transform for Object Recognition[J]. Computer Vision and Image Understanding, 2004, 96(1): 1-16.

9 Du M T. An Improved Generalized Hough Transform for the Recognition of Overlapping Objects[J]. Image and Vision Computing, 1997, 15(12): 877-888.