DOI:10.16262/j.cnki.1000-8217.2011.02.012

·基金纵横·

基于模式识别方法的国家自然科学基金资助重要影响因素分析

郭 松[1] 李恩中[2]

(1 哈尔滨医科大学，哈尔滨 150081；2 国家自然科学基金委员会医学科学部，北京 100085)

1 引言

国家自然科学基金委员会(以下简称自然科学基金委)每年通过发布指南指导申报方向，科技工作者通过申请书的形式进行申请，评议专家也是根据申请书进行评审并给出评议意见，因此申请书的质量是决定能否获得资助的重要依据，尤其是面上项目、青年科学基金以及地区科学基金等项目。

关于如何写好科学基金申请书，已有很多颇有见地的论文可查，本文利用前者没有触及的模式识别方法中的决策森林方法，通过对2009年原生命科学部生物医学工程学科获得与未获得资助的面上和青年科学基金项目的申请书分别进行分析，找出了其中的关键影响因素和一般因素，希望为科研工作者在撰写科学基金申请书时提供参考。

一份申请书能否获得国家自然科学基金资助的影响因素(也可称之为特征)很多，如何从众多的特征中筛选出申请书能否获得最终资助的关键因素，在模式识别中属于有监督学习方法中的特征选择问题。特征选择就是利用机器学习和数据挖掘算法从原始特征中删除那些对分类意义不大的特征，即冗余特征，消除随机干扰因素的影响，最终挑选出一些最具代表性的、对要研究的分类问题具有显著鉴别力的重要特征。目前特征选择方法主要分为3类：过滤法、缠绕法和嵌入法。过滤法是机器学习中用来进行特征选择的最早方法，所有的过滤法都基于数据本身的内在结构信息而不依赖于机器学习算法对特征子集的评价，适合较大的数据集，如 t 检验、非参数得分[1]等。缠绕法依赖于特定分类器的评价指标，其将分类算法嵌入特征选择过程中，是以达到最大分类准确率为引导的一类特征选择方法，计算量大，适合较小的数据集，如遗传算法——支持向量机耦合方法[2]等。过滤法比缠绕法计算复杂度低，速度快，但是过滤法在选择特征的时候，由于不涉及特定的分类器，所以很难确定到底选择由多少个特征组成的特征子集才是最优，而且选择出的特征子集的分类效果依赖于最终使用的分类器，由于与分类器的决策机制脱离，对给定的分类器，由过滤方法选择的特征一般不能使分类器达到最大的分类准确率。而在缠绕法中，由于特征选择的结果靠分类器来评价，选择出的特征与分类器的决策机制能够较好地耦合，从而可以使得分类器的分类准确率达到最大[3]。最后一类为嵌入法，利用具有分类的功能同时具有特征选择功能分类器算法，在分类的过程中，自动进行特征选择，即分类和特征选择并行，例如CART[4]，C4.5[5]以及决策森林方法[6]等。本文就是利用的后者。

2 材料和方法

2.1 数据描述和预处理

为了确定国家自然科学基金申请书最终能否获得资助的诸多因素，本文对原生命科学部生物医学工程学科面上项目和青年项目进行了分别研究。其中面上获得资助项目191份，未获得资助项目894份；青年基金获得资助项目106份，未获得资助项目363份。首先从每一份申请书中提取共同的41项特征，然后将所有的描述性特征都进行数值化。其中，学历特征按照博士、硕士和学士的等级编码成3、2、1；职称则是按照教授、副教授和讲师的等级编码成3、2、1；依托单位则根据2009年中国网大公布的大学和科研机构的排名将其编码到1至100区间内的整数；在研究性质中，应用基础研究为1，基础研究为2；前3申请人专业结构则根据所涉及的专业数量编码为3、2、1；工作条件等级中的A、B、C分

本文于2011年1月10日收到.

别对应数值为 3、2、1。此外，由于部分申请书的某些特征值存在缺失，本文将缺失数据较多的样本进行了删除，最终获得了面上获得资助项目 191 份，面上项目未获得资助项目 892 份；青年基金项目获得资助项目 106 份，青年基金项目未获得资助项目 343 份。

这样就获得了两套包含两类样本的数据集（面上项目获得与未获得资助，青年基金项目获得与未获得资助），它们可以表示成两个矩阵，维数分别为 1083×42 和 449×42，其中前 41 列表示的是样本对应的特征值，最后一列表示的是样本的类别（1 表示获得资助项目，2 表示未获得资助项目）。然后，在两套数据集中分别进行特征选择，确定出国家自然科学基金面上项目和青年项目能否获得资助的关键影响因素。由于该数据中 41 个特征的量纲存在较大差别，因此，在进行特征选择之前，必须对这两套数据进行标准化处理。

这里采用了均值为 0、标准差为 1 的方法对两套数据进行标准化，即：标准化以后的数据都满足均值为 0、标准差为 1，具体公式如下：

$$x'(g,s)=\frac{x(g,s)-\mu(g)}{\sigma(g)},$$

其中 $x(g,s)$ 是第 g 个特征在第 s 个项目中的取值，$\mu(g)$ 和 $\sigma(g)$ 分别表示特征 g 在所有项目中的均值和方差。

2.2 决策森林的构建

为了构建决策森林，首先选用 n 倍交叉证实[7]的方法产生大量的训练集和检验集，首先将每类样本随机分为近似的 n 等份，然后分别从每类样本中选取一份组合构成检验集，余下的所有样本构成训练集，这样，一共有 n^2 种不同的组合，即构造了 n^2 种不同的训练集（L_d）和检验集（T_d）对，这种方法既能使每个样本都参与训练，又保证了原始样本集和训练集中各类样本比率的一致性。

在每个训练集（L_d）上创建一棵递归决策树，本文采用 Gini 差异性指标（代价函数）作为节点的杂质函数。特征选择的过程就是在每个非叶子节点寻找一个最好的特征，使得在正在分叉的节点上，杂质的减少量最大，直到树的增长停止为止。在所有的训练集上重复以上过程，就构建了一个包含 n^2 棵决策树的决策森林。然后，用相应的检验集（T_d）来评价决策树的分类效能，在此，选用正确率（acc）作为评价指标，其计算公式为：

$$\mathrm{acc}=\frac{\mathrm{TP}+\mathrm{TF}}{\mathrm{TP}+\mathrm{NP}+\mathrm{TF}+\mathrm{NF}},$$

其中 TP 为真阳性数（预测和真实都是阳性样本的数目），TF 真阴性数（预测和真实都是阴性样本的数目），NP 假阳性数（预测是阳性样本而真实是阴性样本的数目），NF 假阴性数（预测是阴性样本而真实是阳性样本的数目）。acc 值越高，说明该决策树对样本集的分类效能越好。

在决策树的每一个分类层面上都有一个特征来对样本空间进行划分，这些特征认为是与国家自然科学基金能否获得资助相关的重要因素，因此，每一棵决策树都对应着一个特征子集（G_d），这样就得到了决策森林$\{G_1,\cdots,G_d,\cdots,G_m\}$。然后，综合集成众多的特征选择器得到的特征子集，由各特征被选择的强度（或投票得分）FV 值决定最终的最优特征子集。对每个特征 g_k 可定义并计算各特征的被选择强度：

$$\mathrm{FV}(g_k)=F(G_1,\cdots,G_d,\cdots,G_m)=\frac{\sum_d w_d I(g_k,G_d)}{\sum_d w_d},$$

其中 $I(g_k,G_d)$ 是一个指示函数，当 $g_k\in G_d$，$I(g_k,G_d)=1$；否则，$I(g_k,G_d)=0$，权 w_d 为与基于 G_d 所建立的分类器的分类效能相联系的指标，例如，可取 $w_d=\mathrm{acc}_d$（基于 G_d 所建立分类器的 acc 值）。

3 结果

本文首先采用 5 倍交叉证实的方法对原始样本集进行随机划分，这样，一次交叉证实便能够产生 25 个训练集和检验集对，然后，在每个训练集上构建决策树分类器，并利用检验集进行分类效能的检验，由于一次交叉证实只对原始样本集进行了一次随机划分，所以，重复随机划分 20 次，一共得到了 500 个训练集和检验集对，每对都能得到一个决策树分类器及分类效能，每棵决策树又对应着一个特征子集，最后，利用 FV 值将 500 个特征集合中的因素整合起来，从而获得了各个因素与面上或青年项目获得资助的关联程度。表 1（图 1）和表 2（图 2）分别给出了在面上和青年项目中各因素与获得资助的关联程度。

表 1 各因素与面上项目获得资助的关联强度

特征 ID	特征名称	出现次数	FV 值
22	SCI	7424	14.848
33	已完成的厅局级课题数	6695	13.39
21	近 5 年发表文章	6485	12.97
15	英文参考文献数	6022	12.044
6	申请金额	5350	10.7
32	已完成的省级课题数	5150	10.3
34	已完成的校级课题数	4932	9.864

（续表）

特征 ID	特征名称	出现次数	FV 值	特征 ID	特征名称	出现次数	FV 值
16	中文参考文献数	4890	9.78	39	正在承担的厅局级课题数	3066	6.132
29	已完成的课题数	4704	9.408	30	已完成的国际课题数	2887	5.774
11	博士后人数	4515	9.03	7	总人数	2838	5.676
31	已完成的国家级课题数	4413	8.826	17	拟解决关键问题	2743	5.486
35	正在承担的课题数	4412	8.824	8	高级人数	2719	5.438
25	国家奖数	4037	8.074	9	中级人数	2665	5.33
13	研究生人数	3907	7.814	27	其他奖数	2243	4.486
1	出生年份	3860	7.72	19	创新点	2162	4.324
37	正在承担的国家级课题数	3806	7.612	36	正在承担的国际课题数	2116	4.232
20	重点实验室情况	3797	7.594	26	省部级奖数	1842	3.684
12	博士人数	3626	7.252	41	实验室评级	1367	2.734
10	初级人数	3463	6.926	28	社会兼职	1130	2.26
38	正在承担的省级课题数	3414	6.828	4	依托单位	972	1.944
5	研究性质	3338	6.676	23	前 3 申请人的专业结构	810	1.62
14	参加单位数	3177	6.354	24	获奖情况	718	1.436
40	正在承担的校级课题数	3112	6.224	3	职称	386	0.772
18	可行性分析	3092	6.184	2	学历	327	0.654

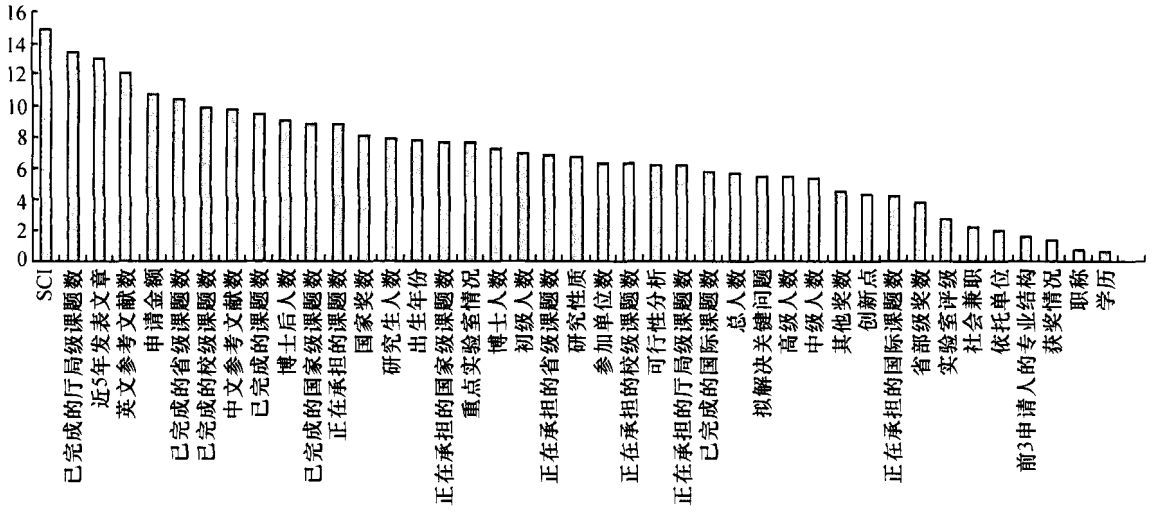

图 1　面上项目中各因素在决策森林特征选择中的 FV 值

表 2　各因素与青年项目获得资助的关联强度

特征 ID	特征名称	出现次数	FV 值	特征 ID	特征名称	出现次数	FV 值
22	SCI	3546	7.092	24	获奖情况	1570	3.14
16	中文参考文献数	3040	6.08	13	研究生人数	1568	3.136
29	已完成的课题数	2788	5.576	1	出生年份	1525	3.05
15	英文参考文献数	2703	5.406	8	高级人数	1474	2.948
21	近 5 年发表文章	2396	4.792	10	初级人数	1415	2.83
35	正在承担的课题数	2330	4.66	17	拟解决关键问题	1376	2.752
11	博士后人数	2316	4.632	18	可行性分析	1223	2.446
12	博士人数	2180	4.36	7	总人数	1209	2.418
31	已完成的国家级课题数	2067	4.134	26	省部级奖数	1147	2.294
6	申请金额	2034	4.068	25	国家奖数	1007	2.014
38	正在承担的省级课题数	1994	3.988	14	参加单位数	983	1.966
40	正在承担的校级课题数	1898	3.796	5	研究性质	851	1.702
3	职称	1866	3.732	27	其他奖	847	1.694
32	已完成的省级课题数	1780	3.56	39	正在承担的厅局级课题数	689	1.378
28	社会兼职	1764	3.528	4	依托单位	582	1.164
34	已完成的校级课题数	1704	3.408	41	实验室评级	537	1.074
37	正在承担的国家级课题数	1680	3.36	30	已完成的国际课题数	500	1
20	重点实验室情况	1662	3.324	36	正在承担的国际课题数	500	1
9	中级人数	1654	3.308	23	前 3 申请人的专业结构	382	0.764
19	创新点	1633	3.266	2	学历	205	0.41
33	已完成的厅局级课题数	1615	3.23				

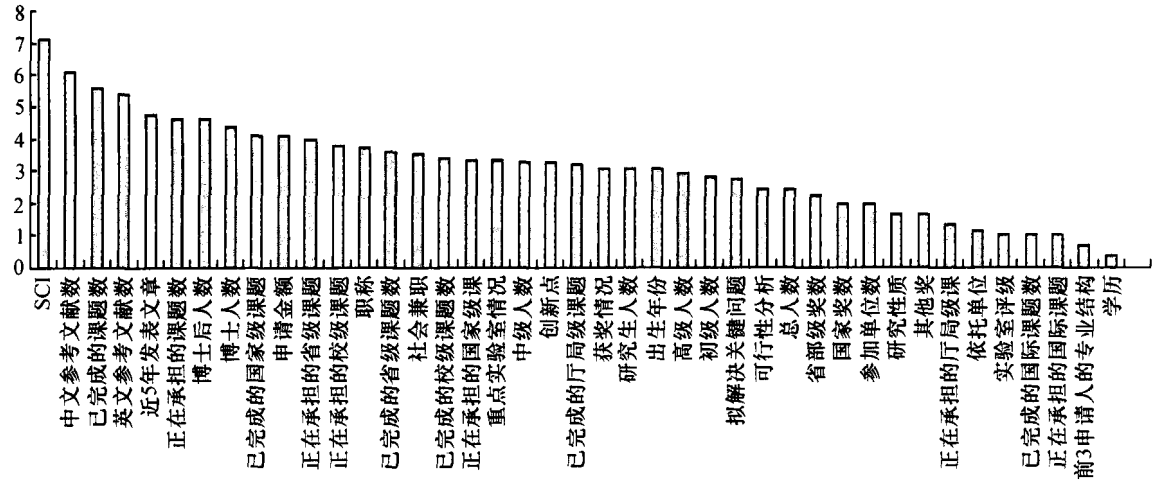

图 2 青年项目中各因素在决策森林特征选择中的 FV 值

4 讨论

结合 FV 值对各因素与项目获得资助的关系进行分析，可以得到以下几点认识：(1) SCI 文章排在第 1 位，其 FV 值最高，同时近 5 年发表的其他文章的 FV 值也很高，说明获得资助与申请人前期研究结果的发表有重要的相关性，是影响申请书获得资助的重要因素；(2) 完成的课题按照厅局级、省级、校级和国家级的顺序，各个级别的 FV 值区分度不大，说明在前期研究工作中要有课题作依托，通过前期课题研究取得一定突破，并通过文章的形式体现研究结果，再对研究结果中的新进展、创新点进行规划，重新开始研究过程；(3) 正在承担的课题排列顺序为国家级、省级、校级和厅局级，区分度不大也对申请书的获得资助起到重要作用；(4) 中、英文参考文献，反映作者的科学态度和求实精神，也表明作者对他人成果的尊重，对某一科学技术领域研究的深度和广度的了解情况，体现了本研究领域前沿和最新进展，在申请书的评审过程中具有重要作用；(5) 项目组成员的学历、职称的比例是否合理对获得资助有一定的影响；(6) 依托单位及重点实验室是完成课题与否的重要依托，具有一定作用；(7) 申请金额虽然 FV 值较高，但每年申请指南有一定要求，只要认真阅读指南都会填写正确；(8) 针对课题内容可行性分析、拟解决的关键问题、创新点等有一定影响，在结果中 FV 值不是很高，可能是在量化过程中的标准所致；(9) 申请人获得奖励的情况、社会兼职影响较小，说明面上项目评审不是人才评审；(10) 申请人的职称和学历影响较小，说明只要符合国家自然科学基金项目的申请人要求都是可以获得资助的对象，评审时区分度不大；(11) 青年基金中，申请人的职称、社会兼职、课题组中级人数、获奖情况 FV 值相对靠前，对申请书获得资助有重要影响，说明青年科学基金的支持青年人才的特殊性，符合国家自然科学基金的资助方向。

5 结论

综上所述，可得以下结论：(1) 面上项目申请书中 SCI 文章、已完成和承担的课题、中英文参考文献对申请书的获得资助起着重要作用；(2) 项目组成员的学历、职称的比例、依托单位及重点实验室是否合理对获得资助有一定的影响；(3) 课题内容可行性分析、拟解决的关键问题、创新点等有一定影响；(4) 申请人获得奖励的情况、社会兼职影响不大；(5) 青年科学基金与面上项目相比，申请人的职称、社会兼职、课题组中级人数、获奖情况对申请书获得资助有重要影响。

参 考 文 献

[1] Park P J, Pagano M, Bonetti M. A Nonparametric Scoring Algorithm for Identifying Informative Genes from Microarray Data. Pacific Symposium on Biocomputing, Mauna Lani, Hawaii USA, June,2001, 52—63.

[2] Li L, Jiang W, Li X et al. A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. Genomics, 2005, 85(1): 16—23.

[3] Li L, Li X, Guo Z et al. Efficiency of two filter feature gene selection algorithms. Life Science Research, 2003, 7(4): 396—372.

[4] Cardie C. Using decision trees to improve case-based learning. Proceedings of the Tenth International Conference on Machine Learning, 1993. Morgan Kaufman. pp. 25—32.

[5] Quinlan J. Induction of decision trees. Machine Learning, 1986. 1:81—106.

[6] Li X, Rao S Q, Wang Y D et al. Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling. Nucleic Acids Research, 2004, 32(9): 2685—2694.

[7] Breiman L. Bagging predictors. Machine Learning, 1996, 24: 123—140.

(转第 110 页)

[5] National Science Board. Overview of Approaches for Identifying, Reviewing, and Supporting Transformative Research, Working Paper of the National Science Board, September, 2004.
[6] 托马斯·库恩著,李宝恒,纪树立译. 科学革命的结构. 上海:上海科学技术出版社,1980 年.
[7] 史蒂芬·科尔著,林建成,王毅译. 科学的制造. 上海:上海人民出版社,2001 年.
[8] National Science Board. Summary of National Science Board Workshop: Identifying, Reviewing and Funding Transformative Research, September, 2004.
[9] Walter E. Stumph. "Peer" Review, Science, 22 February, 1980, 207: 822—823. 转引自 Daryl E Chubin and Edward J Hackett. Peerless Science: Peer Review and U. S. Science Policy. New York: State University of New York Press, 1990.
[10] Alexander A Berezin. Discouragement of innovation by over-competitive research funding. Interdisciplinary Science Reviews, 2001, 26(2): 97—102.
[11] National Science Board. Enhancing Support of Transformative Research at the National Science Foundation. Draft for Public Comment, February 8, 2007.
[12] National Science Foundation. The Second Annual Report of the National Science Foundation: Fiscal Year 1952. U S Government Printing Office, Washington 25, D C, 1952i.
[13] National Academy of Public Administration. National Science Foundation: Governance and Management for the Future, April 2004.
[14] R K 默顿著,鲁旭东,林聚任译. 科学社会学. 北京:商务印书馆,2003.
[15] National Science Board. Report of the National Science Board on the National Science Foundation's Merit Review System, NSB-05-119, September 30, 2005.
[16] National Science Board. Enhancing Support of Transformative Research at the National Science Foundation, NSB-0732, May 7, 2007.
[17] National Science Board. FY 2009 Report on the NSF's Merit Review Process, NSB-1027, May 2010.

PUBLIC FUNDING AND INNOVATIVE RESEARCH —An Analysis on the Policy of Supporting Transformative Research at National Science Foundation of U. S.

Gong Xu
(*Policy Bureau, National Natural Science Foundation of China, Beijing* 100085)

Abstract Using the concept of Kuhn's "paradigm", the paper defines two kinds of innovation research, i. e., accumulative progress and revolutionary breakthrough, and points out the inevitable dilemma that a research funding agency has to face when using its peer review system, which aims to seek for consensus, to identify innovative research proposals, which usually get contradictive comments from reviewers. Based on the analysis on the policy changes from setting up "Small Grant for Exploratory Research" to supporting transformative research at National Science Foundation of U. S., it suggests that a funding agency should not only improve its peer review system to identify potentially transformative research, but also encourage scientists to submit their innovative ideas, and develop more measurements, such as multi-disciplinary joint funding project and workshops called "Ideas Factory Sandpit", to promote the support of transformative research.

Key words NSF, transformative research, research policy, peer review

(接第 128 页)

ANALYSIS OF IMPORTANT FACTORS FOR FUNDING FROM THE NATIONAL NATURAL SCIENCE FOUNDATION OF CHINA BY THE METHOD BASED ON PATTERN RECOGNITION

Guo Song[1] Li Enzhong[2]
(1 *Harbin Medical University, Harbin* 150081; 2 *Department of Health Sciences, NSFC, Beijing* 100085)