文章编号: 1001-0920(2004)08-0927-04

# 基于 SVM 的中文文本分类反馈学习技术的研究

孙晋文, 肖建国
(北京大学 计算机科学技术研究所, 北京 100871)

**摘 要**: 基于相关反馈技术的基本原理, 以 SVM 分类方法为基础, 研究了基于 SVM 的中文文本分类反馈学习技术, 分析了分类处理中反馈学习的主要模式, 给出了基于 SVM 文本分类反馈学习的具体实现方法, 并进行了相应的实验验证. 实验结果表明, 反馈学习具有明显提高 SVM 分类性能的能力.

**关键词**: 中文文本; 支持向量机; 反馈; 学习
**中图分类号**: TP18 **文献标识码**: A

# Study on feedback learning of SVM-based chinese text classification

*SUN Jin-wen, XIAO Jian-guo*
(Institute of Computer Science and Technology, Peking University, Beijing 100871, China. Correspondent, SUN Jin-wen, E-mail: sunjinwen@sina.com)

**Abstract**: Based on the principle of feedback learning, the feedback ways of SVM-based chinese text classification are discussed. The different patterns in feedback processing are analyzed. The detail feedback algorithm on SVM-based chinese text classification is presented. The experimental results show that the feedback learning can greatly improve the performance of SVM classification.

**Key words**: chinese text; SVM; feedback; learning

## 1 引 言

数字化信息资源极大丰富, 使得信息分类技术成为知识挖掘的重要内容. 对于文本自动分类技术的研究, 支持向量机[1~4](SVM)技术已成为分类研究的重要方法.

分类技术是模式识别技术的一个重要应用, 模型的学习是其核心内容, 各种不同分类方法的研究, 在很大程度上是为了从不同的方面更好地提高模型的学习效果. 在模型的改善中, 反馈技术[5~9]是一种重要的研究方法, 它通过人机交互的方式将模型的输出通过一定方式返回到输入中, 以改善模型的性能. 大量研究成果表明了反馈对改善检索的作用. 然而, 作为信息挖掘重要应用的文本分类技术, 反馈方法在此方面的研究却很少. 为此, 本文以 SVM 分类方法为基础, 着重探讨反馈技术在文本分类中的有效性, 探讨了反馈技术在基于 SVM 分类中的应用类型, 分析了基于 SVM 中文文本反馈学习技术的具体实现方法, 并进行了相关实验. 实验结果验证了其性能的有效性.

## 2 基于 SVM 的中文文本分类反馈学习

支持向量机作为一种新的通用的机器学习方法, 已成为当今研究的重点, 其具体的理论在此不再赘述.

### 2.1 问题描述

为表述方便, 本文作如下规定:

**定义 1** 文档集合分为 3 类: 训练文档集合 $E$,

**收稿日期**: 2003-07-31; **修回日期**: 2003-11-18.

**作者简介**: 孙晋文(1972—), 男, 山东枣庄人, 博士后, 从事人工智能、数据挖掘等研究; 肖建国(1957—), 男, 辽宁鞍山人, 教授, 博士生导师, 从事信息处理、网络系统集成的研究.

测试文档集合 $T$, 待分类文档集合 $U$. $\Omega$ 为训练后产生的分类器, $S$ 为分类器 $\Omega$ 中的支持向量集, $R$ 为经分类器 $\Omega$ 分类处理后的结果集, 并用 $R^E, R^T, R^U$ 分别表示集合 $E, T, U$ 的分类结果集; $F$ 为人工交互后用于反馈学习的文档集合, $d = \{(t_1, w_1), \cdots, (t_m, w_m)\}$ 为文档向量, 用 $d^{sv}$ 和 $d^{fb}$ 分别代表 $S$ 中的支持向量和 $F$ 中的反馈文档向量; $C = \{c_i \mid i = 1, \cdots, m\}$ 为预定义的类别, $m$ 为类别数, 并用 $C^{init}, C^{auto}, C^{fb}$ 分别表示文档的原始分类类别、自动分类类别、反馈分类类别集合; $P = \{p_1, p_2, p_3\}$ 作为文档集合类型标记, 分别表示文档为训练样本、测试样本、待分类样本. 则

$$E = \{\langle d_i, c_i^{init}, p_1\rangle \mid i = 1, \cdots, n^E\},$$
$$T = \{\langle d_i, c_i^{init}, p_2\rangle \mid i = 1, \cdots, n^T\}.$$
$$U = \{\langle d_i, p_3\rangle \mid i = 1, \cdots, n^U\},$$
$$S = \{d_i^{sv} \mid i = 1, \cdots, n^S\}.$$

其中: $n^E, n^T, n^U, n^S$ 分别为集合 $E, T, U, S$ 的大小; $R = \{R^E, R^T, R^U\}$, $R^E = \langle E, C^{auto}\rangle$, $R^T = \langle T, C^{auto}\rangle$, $R^U = \langle U, C^{auto}\rangle$; $F = \{\langle d_i^{fb}, c_i^{fb}, p_j\rangle \mid d_i^{fb} \in R, p_j \in P, i = 1, \cdots, n^F, j = 1, 2, 3\}$; $n^F$ 为集合 $F$ 的大小.

对于基于 SVM 分类而言, 寻找位于不同类别边界处的点是分类模型学习的主要任务. 模型的分类处理主要与分类边界处的支持向量有关, 而其他文档对分类处理不起作用. 对基于 SVM 分类的反馈学习而言, 经系统分类后证明是正确的文档, 表明原分类模型已包含该文档的相关分类信息, 因此它对反馈学习没有价值. 而分类错误或不能分类的文档, 则说明含有原模型中所不包含的分类信息, 是进行反馈学习的重点. 因此, 基于支持向量机分类反馈学习的重点是针对误分类文档或不能分类文档, 而这些文档往往是整个分类结果集合的一小部分. 从理论上讲, 该学习方法具有反馈样本少的优点.

### 2.2 分类反馈学习的模式类型

对于分类处理, 根据文档集合的不同可分为训练集、测试集和待分类集. 不同的集合在反馈处理时有一定的不同, 因此反馈学习也可划分为以上 3 种类型. 另外, 根据反馈文档与分类模型中支持向量关系的不同, 也可以将反馈学习分为封闭反馈学习和开放反馈学习. 其中封闭反馈学习是指对训练集的反馈, 而对测试集、待分类文档集的反馈则为开放反馈学习, 相关处理方法如下:

1) 训练集的反馈: 在训练集 $E$ 训练完成后, 通过获得的分类模型 $\Omega$ 对 $E$ 进行封闭测试, 产生分类结果集 $R^E$. 对其中的错误分类文档集 $R^E_{error}$ 进行反馈处理, $R^E_{error} \subset R^E$, 一般 $R^E_{error} \ll R^E$. 若 $R^E_{error} = \varnothing$, 则结束; 否则, 对于 $(d_i, c_i^{init}, c_i^{auto}) \in R^E_{error}$, 通过人工交互, 若原文档分类类别 $c_i^{init}$ 不准确, 且其类别相关度高, 则由人工给出其正确的人工分类类别 $c_i^{fb}$, 并保存到反馈集合 $F$ 中. 对于训练文档集的反馈处理, 可用于发现训练文档中的噪声数据(原始分类不正确的文档), 从而纠正训练文档中的错误数据信息, 提高原始分类模型的质量.

2) 测试集的反馈: 测试集的反馈处理流程与训练集既基本一致, 又具有很大不同, 测试集可以认为是具有人工分类信息的待分类文档. 对于由训练集合 $E$ 产生的分类器 $\Omega$, 选取测试集合 $T$, 由分类器 $\Omega$ 进行分类处理, 产生分类结果集 $R^T$, 其中错误分类集为 $R^T_{error}$, 对 $R^T_{error}$ 进行人工反馈. 对于 $(d_i, c_i^{init}, c_i^{auto}) \in R^T_{error}$, 因为 $c_i^{init} \neq c_i^{auto}$, 若经过人工反馈处理后, 发现 $c_t^{init}$ 正确而 $c_i^{auto}$ 错误, 同时该文档 $d_i$ 具有较高的类别相关性, 则将其保存到反馈集合 $F$ 中, 并将 $c_i^{init}$ 赋值给 $c_i^{fb}$. 通过对测试集的分类反馈处理, 可以弥补训练文档所产生分类模型的不足, 进一步提高分类模型的分类性能.

3) 待分类文档集的反馈: 待分类文档集是实际分类应用中需要处理的大量无类别信息的文档, 是分类模型在应用中的主要对象. 对于已有的分类器 $\Omega$, 待分类文档集 $U$ 经分类处理后获得的结果集为 $R^U$, 反馈处理需要对整个 $R^U$ 进行. 对于结果集中自动分类 $c_i^{auto}$ 错误, 且类别相关度高的文档, 给出其相应反馈类别 $c_i^{fb}$, 并将其存入反馈集 $F$ 中, 用于反馈训练学习. 若系统可以给出不能分类文档集, 不妨记作 $R^U_{un}$, 则也可重点对 $R^U_{un}$ 进行反馈处理.

### 2.3 反馈学习中的支持向量优化

对于支持向量机分类而言, 通过二次优化训练处理, 获得位于分类边界处由数据点组成的集合, 用于分类处理. 这些数据点, 即支持向量, 包括位于分类边界上的点和分类边界间不能正确分类的数据点. 对于支持向量集 $S$, 一般是训练集合 $E$ 的一个比较小的子集, 通常 $S \ll E$. 通过对支持向量机的原理分析可知, 支持向量包含了所有训练文档集合 $E$ 中的有用分类信息. 因此, 对于支持向量机而言, 具有良好的增量训练学习的特性, 而对于反馈学习, 则可以充分利用其增量学习的优点, 对反馈文档集 $F$ 和模型 $\Omega$ 中支持向量集 $S$ 重新进行优化选择, 获取新的分类模型 $\Omega'$ 及支持向量集 $S'$. 算法相对比较简单, 具体如下:

训练集合 $E$, 经训练产生分类器 $\Omega$, $S$ 为分类模

型中的支持向量集, 反馈文档集为 $F$. 将 $S$ 与 $F$ 的并集 $E'$, $E' \leftarrow S \cup F$, 进行优化训练, 产生新的分类模型 $\Omega'$ 和新的支持向量集 $S'$.

### 2.4　训练文档反馈的支持向量消重

主要针对训练文档集 $E$ 的反馈处理. 由于反馈文档进行反馈学习时, 反馈文档集需要与原模型的支持向量集一起进行优化训练, 而训练文档的反馈集合主要由具有不正确原始分类的训练文档组成. 这些文档有可能已成为原分类模型 $\Omega$ 的支持向量, 若仍留在原支持向量集合 $S$ 中, 则会影响模型的效果. 因此, 在进行训练文档的反馈学习过程中, 反馈训练前需首先对该类反馈文档进行支持向量的检测, 进行消重处理, 以消除对分类性能的影响. 为更好地进行支持向量的相似消重处理, 定义 $\theta_{sim}$ 为文档相似度系数, 作为进行支持向量消重的阈值, 具体算法如下:

Step 1: 对反馈文档集合 $F$ 中的任一反馈文档进行文档类型判断, 若该文档为训练文档, 则其反馈文档向量为 $d_j^{fb}(t_i, w_i)$, $j = 1, \cdots, n^F$;

Step 2: 读取支持向量集合 $S$ 的支持向量 $d_k^{sv}(t_i, w_i)$, $k = 1, \cdots, l$; $l$ 为向量集合 $S$ 中支持向量数;

Step 3: 对于任一训练反馈文档向量 $d_j^{fb}$ 和支持向量 $d_k^{sv}$, 计算其向量相似度

$$\mathrm{sim}(d_j^{fb}, d_k^{sv}) = \frac{d_j^{fb} d_k^{sv}}{\| d_j^{fb} \| \| d_k^{sv} \|};$$

Step 4: 进行文档向量相似度判断, 若 $\mathrm{sim}(d_j^{fb}, d_k^{sv}) \geq \theta_{sim}$, 则判定该两篇文档相同, 将 $d_k^{sv}$ 从支持向量集合 $S$ 中删除.

### 2.5　反馈处理算法流程

针对以上处理, 具体反馈处理算法描述如下:

Step 1: 对分类结果集 $R$ 中任一篇文档 $d_i \in R$ 进行反馈处理, 若文档属于训练文档或测试文档, 则可只对 $R$ 中的错误分类结果集 $R_{error}$ 进行处理, 根据文档的内容特征对需要进行反馈处理的文档, 给出相应的人工反馈类别 $c_i^{fb}$, 存入反馈文档集 $F$;

Step 2: 所有文档反馈处理完毕后, 开始进行反馈训练学习, 从反馈集 $F$ 中读取任一文档, 生成相应文档向量 $d_j^{fb}(t_i, w_i)$, $j = 1, \cdots, n^F$;

Step 3: 从原分类模型 $\Omega$ 的支持向量集合 $S$ 中读取支持向量 $d_k^{sv}(t_i, w_i)$, $k = 1, \cdots, l$;

Step 4: 对集合 $F$ 的反馈文档进行文档类型判断, 若文档 $d_j^{fb}(t_i, w_i)$ 属于训练文档的反馈文档, 则对其与 $d_k^{sv}(t_i, w_i)$ 进行支持向量消重处理;

Step 5: 重复 Step 4, 直到所有训练反馈文档支持向量消重完毕, 获得消重后的支持向量集 $S'$;

Step 6: 取反馈学习文档集 $F$ 与支持向量 $S'$ 的并集作为训练文档集, 进行支持向量机的重新优化训练.

## 3　实验分析

为验证基于支持向量机分类反馈训练学习的效果, 本文进行了如下实验: 数据为人民网 2001 年和 2002 年新闻分类语料, 分为体育、信息技术、军事、文娱、科教、环保、经济共 7 大类别, 共 10 000 多篇. 根据反馈类型不同, 分别进行了训练文档和非训练文档的反馈学习实验. 非训练文档应包括测试文档及待分类文档, 在本实验中都用测试文档代替. 下面具体介绍相关的实验情况.

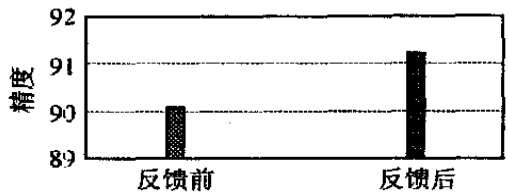
反馈实验结果如表 1 和图 1 所示. 为更好地验证反馈的效果, 训练文档为具有 10% 干扰数据的数据集合.

通过反馈实验可以看出:

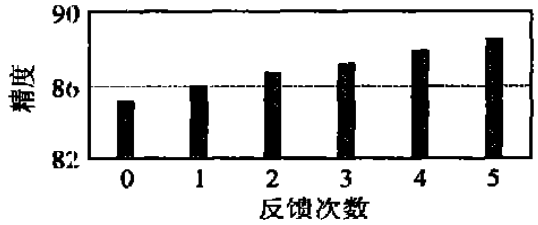
1) 反馈学习对分类性能有明显的提高作用, 实验中反馈前后系统分类精度分别从 90.095% 提高到 91.19%, 分类精度提高了近 1 个百分点; 而在非

表 1　分类反馈实验结果

| 分类精度/% | | 体育 | 信息技术 | 军事 | 文娱 | 科教 | 环保 | 经济 | 平均 |
|---|---|---|---|---|---|---|---|---|---|
| 训练文档 | 反馈前 | 96.67 | 87.67 | 94.67 | 90.67 | 81.33 | 92 | 87.67 | 90.095 |
| | 反馈后 | 97 | 88.67 | 96.33 | 92 | 85 | 91 | 88.33 | 91.19 |
| 非训练文档 | 初始训练集 | 96.5 | 77.5 | 90 | 88 | 81 | 85.5 | 77.5 | 85.143 |
| | 反馈 1 | 96.5 | 79 | 93 | 90 | 82.5 | 85.5 | 76 | 86.0 |
| | 反馈 2 | 97.5 | 83 | 92.5 | 86.5 | 81.5 | 86 | 80 | 86.714 |
| | 反馈 3 | 97 | 82.5 | 92 | 87 | 83 | 86.5 | 82 | 87.143 |
| | 反馈 4 | 97 | 87.5 | 92.5 | 90 | 80 | 88 | 80 | 87.857 |
| | 反馈 5 | 97.5 | 89.5 | 92.5 | 89 | 82 | 91.5 | 77 | 88.429 |




(a) 训练文档(10% 干扰数据) 反馈学习前后对比




(b) 非训练文档分类反馈数据集实验结果

**图 1 反馈实验结果**

训练文档的反馈实验中, 经 5 次反馈学习后, 系统的分类精度从 85. 143% 提高到 88. 429%, 分别提高了 3～ 4 百分点, 反馈学习效果明显, 而且从非训练文档的实验结果可以看出, 随着反馈的进行, 分类性能逐步提高

2) 反馈训练具有反馈学习数据少的优点

3) 由于反馈学习需要人工给出相应的反馈类别, 在文档本身存在较强兼类特性的情况下, 可能会出现人工分类标准与实际训练样本标准不一致的情况, 造成反馈学习后的某个类别分类性能暂时波动但这不影响反馈学习对整体分类性能的提高, 同时也说明了学习样本质量对分类性能影响的重要性

总之, 从对基于 SVM 的文本分类的研究可以看出, 反馈学习是文本分类的一种有效的学习方法可以通过较小的反馈文档数量, 实现较大的分类性能提高, 具有反馈样本少, 效果提高明显的优点 因此, 该算法是进行文本分类研究与应用的有效方法

**参考文献**(References):

[1] Vapnik V. *The Nature of Statistical Learning Theory* [M]. New York: Springer-Verlag, 1995.

[2] Osuna E, Freund R, Girosi T. Training support vector machines: An application to face detection[A]. *Proc of the IEEE Int Conf on Computer Vision and Pattern Recognition*[C]. New York, 1997. 130-137.

[3] Joachims T. Text categorization with support vector machines: Learning with many relevant features[A]. *Proc of the European Conf on Machine Learning*[C]. Berlin: Springer, 1998. 137-142.

[4] Yang Yiming, Liu Xin. A re-examination of text categorization methods[A]. *Proc of ACM SIGIR Conf on Research and Development in Information Retrieval*[C]. Berkeley, 1999. 42-49.

[5] Rocchio J J. Relevance feedback in information retrieval [A]. *The SMART Retrieval System Experiments in Automatic Document Processing* [C]. Ne Jersey: Prentice Hall Inc, 1971. 313-323.

[6] Ide E. Relevance feedback in an automatic document retrieval system[R]. Ithaca, NY: Cornell University, 1969.

[7] Salton G. *The SMART Retrieval System* [M]. Englewood Cliffs N J: Prentice Hall, Inc, 1971.

[8] Cox I J, Miller M L, Omohundro S M, et al. Pichunter: Bayesian relevance feedback for image retrieval system[A]. *Int'l Conf on Pattern Recognition* [C]. Vienna, Austria, 1996. 361-369.

[9] Lee Joon Ho. Combining the evidence of different relevance feedback methods for information retrieval [J]. *Information Processing & Management*, 1998, 34 (6): 681-691.

(上接第 922 页)

[5] Hsu Chih-Chin, Fong I-Kong. Ultimate boundedness control of linear systems with band-bounded nonlinear actuators and additive measurement noise[J]. *Systems & Control Letters*, 2001, 43(4): 329-326.

[6] Truxal J G. *Control Engineer's Handbook*[M]. New York: McGraw-Hill, 1958.

[7] Hsu K C. Variable structure control design for uncertain dynamic systems with sector nonlinearities [J]. *Automatica*, 1998, 34(4): 505-508.

[8] Hsu K C. Decentralized variable-structure model-following control design for interconnected systems with series nonlinearities[J]. *Int J of Systems Science*, 1998, 29(4): 365-372.

[9] Liu F L, Wu H, Ma C G, et al. Decentralized dynamic output feedback control for a class of similar interconnected large-scale systems with nonlinear input [A]. *Proc of World Multi Conf on Systemics, Cybernetics and Informatics*[C]. Orlando, 2001, IX: 142-147.