第 40 卷 第 2 期
Vol.40 No.2

计 算 机 工 程
Computer Engineering

2014 年 2 月
February 2014

• 人工智能及识别技术 •　文章编号：1000—3428(2014)02—0171—04　文献标识码：A　中图分类号：TP18

基于 CPSO 的多目标文本分类投影寻踪

石 松，陈 云

(上海财经大学上海市金融信息技术研究重点实验室，上海 200433)

摘 要：投影寻踪可有效解决文本分类中的维数灾难问题，而投影方向优化是投影寻踪需要解决的关键问题。传统的投影寻踪方法将投影指标优化看作单目标优化问题，会使解的质量受到影响。为此，提出一种基于多目标优化的投影寻踪方法。将类别之间的距离和类别内数据的聚类紧密程度作为 2 个优化目标，并将投影扩展到多维，利用混沌粒子群优化算法寻找最优的投影方向。在常用文本数据集上进行实验，确定最优投影指标及维度，并比较不同分类模型的分类结果，结果表明，使用该方法能有效提高文本分类性能。

关键词：投影寻踪；文本分类；维数灾难；投影指标；多目标优化；混沌粒子群优化算法

Multi-objective Projection Pursuit for Text Categorization Based on CPSO

SHI Song, CHEN Yun

(Shanghai Key Laboratory of Financial Information Technology, Shanghai University of Finance and Economics, Shanghai 200433, China)

【Abstract】Projection pursuit method is increasingly used in text categorization to solve the curse of dimensionality. Traditional projection pursuit method considers the projection index optimization as a single-objective problem rather than a multi-objective one, which will reduce the quality of the solution. To solve this problem, this paper proposes a projection pursuit mehod based on multi-objective optimization. Measures are taken like class difference and difference between the classes as two objectives of pursuit index, the projection pursuit method is extended to multi-dimensional projections, and a Chaotic Particle Swarm Optimization(CPSO) is suggested to find the optimal projection direction. Experiment on commonly used text datasets determines the optimal projection direction and dimensions, and then compares the results of different classification models. The results demonstrate that the proposed method can improve the text categorization performance effectively.

【Key words】projection pursuit; text categorization; curse of dimensionality; projection index; multi-objective optimization; Chaotic Particle Swarm Optimization(CPSO) algorithm

DOI: 10.3969/j.issn.1000-3428.2014.02.037

1 概述

文本分类是文本挖掘的重要领域之一，已经被广泛应用到许多领域，日益为研究者所关注。文本被表示成向量空间模型后，数据空间的维度往往非常高。高维数据具有计算量大、数据稀疏等特征，这些特征导致了维数灾难。因此，在分类之前往往要降维[1]。投影寻踪是处理高维数据，尤其是来自非正态总体的高维数据的一种新型的统计方法[2]。

随着数据挖掘领域的发展，投影寻踪模型被越来越广泛地用来解决分类和聚类中的维数灾难问题。文献[3]为能将投影后所得数据的不同类别分离开，采用在方向投影后的数据的全局误判率作为投影寻踪指标，并应用到 2 个类的分类问题中。文献[2]采用模拟退火算法最大化有监督分类的投影指标。文献[4]采用将投影寻踪应用到洪水灾害分类评价模型中，并使用多种群合作粒子群优化算法(Multi-swarm Cooperative Particle Swarm Optimization, MCPSO)找到最优的投影方向。在文本分类的研究中，文献[5]采用遗传算法优化投影方向；文献[6]采用了粒子群算法和遗传算法的混合算法作为优化算法；文献[7]提出了基于免疫进化算法的投影寻踪模型。

投影方向的优化是投影寻踪需要解决的关键问题。传

基金项目：上海市科学技术委员会基金资助项目(10dz1123500, 10dz1123200, 11ZR1411800)；上海市自然科学基金资助项目(11ZR1411800)；上海财经大学研究生创新基金资助项目(CXJJ-2012-322)

作者简介：石 松(1985－)，男，博士研究生，主研方向：数据挖掘，信息检索，智能算法；陈 云，教授、博士生导师

收稿日期：2012-10-30　**修回日期：**2013-02-21　**E-mail：**shisongjxnu@163.com

统的投影寻踪方法通过函数聚集把投影指标优化看成是单目标优化问题。这种方法每次迭代只能产生一个最优解，没有精英保留机制，这样会使解的质量受到影响。另外，以往研究在将投影寻踪应用到文本分类问题时，投影后的数据集中在一维空间，没有扩展到多维。投影到多维对于研究高维数据的性质十分重要。

针对上述问题，本文将投影指标优化看作多目标优化问题，将类别之间的距离和类别内的聚类紧密程度作为 2 个优化目标，并考虑将数据投影到多维，使用混沌粒子群优化算法(Chaotic Particle Swarm Optimization, CPSO)寻找最优的投影方向。

2 相关概念

2.1 基于投影寻踪的文本分类

投影寻踪的基本思想是通过优化某个函数(投影指标)将高维数据投影到低维(1~3 维)子空间上，从而达到降低维度的目的。

在利用投影寻踪对文本分类的过程中，首先需要根据分类的目标构造用于寻找最优投影方向的投影指标；然后使用优化算法优化该投影指标找到最优的投影方向，将高维数据投影到低维的投影方向上；最后用分类器对低维数据进行分类。

2.2 多目标优化

多目标优化问题是模拟优化包含 2 个或 2 个以上函数的集合 S。这些函数通常用来描述一个解决方案的不同特征。由于这些特征之间会存在冲突，因此不能用一个单一的函数来描述问题的解。但可以得到最优的集合。这个集合通常被称作 Pareto 最优集合[8]。

2.3 混沌粒子群优化算法

粒子群优化(PSO)[9]算法，源于对鸟群捕食行为的研究，从鸟群捕食模型当中得到启示，并用于解决优化问题。

为了避免粒子陷入局部最优，提高粒子群算法的全局搜索能力。研究者提出了基于 Zaslavskii 混沌映射[10]的混沌粒子群算法，将 Zaslavskii 混沌映射生成的随机序列来代替算法中的随机数，提高算法的随机选择性能[11]。Zaslavskii 混沌映射是一个有趣的随机系统，它的表示方法如下：

$$X_{n+1}=(X_n+v+aY_{n+1})\bmod(1) \tag{1}$$

$$Y_{n+1}=\cos(2\pi X_n)+\mathrm{e}^{-r}Y_n \tag{2}$$

其中，n 为循环次数。当 v=400，a=12.6695，r=3 时，X 和 Y 就表现出很好的混沌特性。观察可知，$Y_{n+1}\in[-1.0512,1.0512]$。

在改进过程中，用混沌映射取代了 r_{1j} 和 r_{2j} 的值，改进后的速度更新公式如下：

$$v_{ij}(t+1)=\omega\times v_{ij}(t)+c_1\times CM_1\times(p_{ij}(t)-x_{ij}(t))+ c_2\times CM_2\times(p_{gj}(t)-x_{ij}(t)) \tag{3}$$

其中，$0\leqslant CM_1,CM_2\leqslant 1$ 是基于混沌映射的函数。实验结果表明，混沌映射可以提高 PSO 的遍历性、不规则性和随机性，提高算法的寻优性能。

3 基于 CPSO 的多目标投影寻踪文本分类模型

基于 CPSO 的多目标投影寻踪有监督文本分类过程如下：

(1)选取投影指标；

(2)用多目标粒子群算法优化该投影指标找到最优的投影方向；

(3)用最优的投影方向将高维数据投影到低维空间；

(4)采用常用的文本分类器(本文使用的是台湾大学的 LIBSVM)对低维数据分类。

3.1 投影指标的选取

在投影寻踪的理论和应用研究问题中，投影指标的选取和投影方向的优化是关键性的问题。特别地，投影指标的选择是投影寻踪能否成功的关键。

3.1.1 单目标投影指标

在常用的文本分类投影寻踪指标中，数据集中的每个文本被表示为一个 p 维向量，p 是数据集中特征词的个数，$X_{i,j}$ 表示第 i 类数据集的第 j 篇文本，$i=1,2,\cdots,g$，g 是类别总数，j=1,2,…,n_i，n_i 是类别 i 中的文本总数。$\boldsymbol{\alpha}(\boldsymbol{\alpha}\in\mathbf{R}^p)$ 为投影方向。记 $y_{i,j}=\boldsymbol{\alpha}^{\mathrm{T}}X_{i,j}$ 是 $X_{i,j}$ 投影到 1 维空间后的投影数据。$\overline{y_i}=\sum_{j=1}^{n_i}y_{i,j}/n_i$ 表示投影后第 i 类数据的类中心。$\overline{y_{..}}=\sum_{i=1}^{g}\sum_{j=1}^{n_i}y_{i,j,l}/n$ 表示投影后的全局中心，$n=\sum_{i=1}^{g}n_i$。

常用的文本分类投影寻踪指标的定义如下：

$$D(\boldsymbol{\alpha})=\left[\frac{\sum_{i=1}^{g}|\overline{y}_i-\overline{y}_{..}|^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}|y_{i,j}-\overline{y}_i|^2}\right]^{1/2} \tag{4}$$

其中，$\sum_{i=1}^{g}|\overline{y_i}-\overline{y_{..}}|^2$ 是对不同类别之间的观测值的距离的度量；$\sum_{i=1}^{g}\sum_{j=1}^{n_i}|y_{i,j}-\overline{y_i}|^2$ 是对同一类别内的聚类紧密程度的度量。维数约减通过最大化上述投影指标，找到最优的投影方向 $\boldsymbol{\alpha}$ 来实现。

3.1.2 多维单目标投影指标

为了将文本分类中的数据投影到多维，本文提出了多维单目标投影寻踪指标，令 $A=[\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_k]$ 为到 k 维空间的投影，记 $y_{i,j,l}=\boldsymbol{\alpha}_l^{\mathrm{T}}X_{i,j}$ 是 $X_{i,j}$ 投影到 k 维空间中的第 l 维的投影数据。$\overline{y}_{i\cdot l}=\sum_{j=1}^{n_i}y_{i,j,l}/n_i$ 表示投影后第 i 类数据投影到第 l 维的类中心。$\overline{y}_{\cdot\cdot l}=\sum_{i=1}^{g}\sum_{j=1}^{n_i}y_{i,j,l}/n$ 表示投影到第 l 维的全局中心，$n=\sum_{i=1}^{g}n_i$。k 维空间上的投影指标定义为：

$$D(A)=\left[\frac{\sum_{l=1}^{k}\sum_{i=1}^{g}|\overline{y}_{i\cdot l}-\overline{y}_{\cdot\cdot l}|^2}{\sum_{l=1}^{k}\sum_{i=1}^{g}\sum_{j=1}^{n_i}|y_{i,j,l}-\overline{y}_{i\cdot l}|^2}\right]^{1/2} \tag{5}$$

可以看出，式(4)是式(5)在 k=1 时的特殊情况。维数约减通过最大化上述投影指标，找到最优的投影方向 A 来实现。

3.1.3　多维多目标投影指标

单目标优化问题每次迭代只能产生一个最优解，没有精英保留机制，这样会使解的质量受到影响。因此，本文提出了用于文本分类的多维多目标投影寻踪指标。令 $A=[\alpha_1,\alpha_2,\cdots,\alpha_k]$ 为到 k 维空间的投影，多维多目标投影指标的定义如下：

最大化：

$$f_1(A)=(\sum_{l=1}^{k}\sum_{i=1}^{g}|\bar{y}_{i\cdot l}-\bar{y}_{\cdot\cdot l}|^2)^{1/2} \tag{6}$$

最小化：

$$f_2(A)=(\sum_{l=1}^{k}\sum_{i=1}^{g}\sum_{j=1}^{n_i}|y_{i,j,l}-\bar{y}_{i\cdot l}|^2)^{1/2} \tag{7}$$

维数约减通过优化上述 2 个目标函数，找到最优的投影 A 来实现。

3.1.4　多目标优化投影指标

受多目标优化非劣解集的启发，本文提出了多目标优化投影指标。在多目标优化投影指标中，多目标优化算法通过优化式(6)、式(7)中 k=1 时的情况，得到多个能将高维数据投影到 1 维的非劣解；在非劣解集合中选取最优的 K 个投影方向将数据投影到 K 维低维空间。

3.2　多目标优化过程

多目标优化的实现过程如文献[12]所述。一个外部集合 ER(非劣解集)用于存储目前找到的全局最优位置。该集合初始化为种群中的非劣解集，一旦出现了更好的解，就会更新外部集合。

粒子的最优包括个体最优粒子和群体最优粒子，其中个体最优位置的更新方式是通过比较当前粒子和该粒子局部最优位置的支配关系，如果当前粒子的位置能支配粒子的局部最优位置，则将当前位置作为粒子的局部最优位置。群体最优位置为从非劣解集中随机选择的一个粒子。

迭代完成后，在非劣解集合中选择最优的粒子，即对应单目标值最优的粒子作为最优的投影方向。

3.3　参数设置

本文实验涉及的参数设置如下：最大进化代数为 500；粒子数目为 21；每个粒子的维度为 $K\times$14 206，K 表示投影后数据的维度，14 206 是数据集中特征词的总数；惯性权重 $\omega=0.6$；学习因子 $c_1=c_2=1.7$；粒子群中的粒子位置初始化在[0,1]之间；$V_{\min}=X_{\min}=0$，$V_{\max}=X_{\max}=1$。

4　实验结果与分析

本文使用文本分类研究中常用的数据集 20News Groups(20NG)。在实验过程中，选用了其中 5 个新闻组的数据，分别为 comp.graphics、comp.os.ms-windows.misc、comp.sys.ibm.pc.hardware、comp.sys.mac.hardware 和 comp.windows.x，共计 5 000 篇文档。

在预处理过程中，先去掉文档中的停用词，使用 Porter 算法进行词干化处理。使用 LTC 权重公式计算特征词的权重，为了加速处理并减少统计误差，移去了文档频数小于 3 的单词。预处理后，数据集中的特征词个数为 14 206，这也是原始数据的维度。

实验采用 10-折交叉验证的方法，采用数据集上各类的宏平均(Macro Average)F1 指标和微平均(MicroAverage)F1 指标作为衡量分类模型性能的评价标准。首先比较了将数据投影到不同维度和不同投影指标对文本分类结果的影响；接下来比较了采用不同的分类模型对分类结果的影响。

4.1　最优投影指标及维度的确定

实验比较了采用单目标粒子群算法(PSO)、多目标粒子群算法(MOPSO)和多目标混沌粒子群算法(MOCPSO)作为优化算法，采用不同的投影指标，在不同维度上(1~3)的投影结果，以确定最优的投影方式。实验结果如图 1~图 3 所示。其中，横坐标表示投影的维度，纵坐标表示分类结果的宏平均 F1 指标值和微平均 F1 指标值的比值 λ。

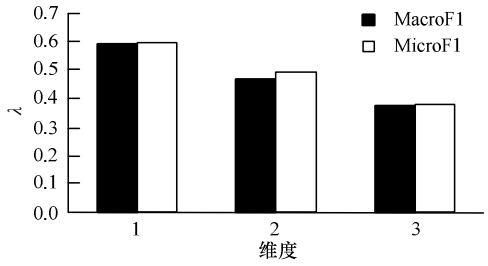

图 1　PSO 在不同维度下的分类结果比较

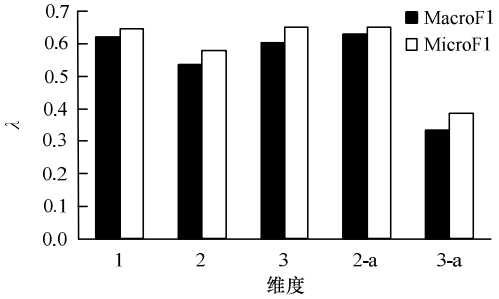

图 2　MOPSO 在不同维度下的分类结果比较

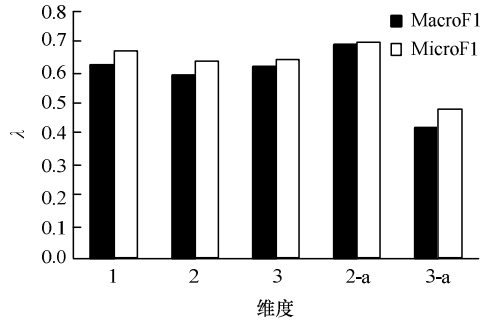

图 3　MOCPSO 在不同维度下的分类结果比较

图 1 给出了单目标粒子群算法(PSO)将数据投影到不同维度下(1~3)后再分类的分类结果的宏平均 F1 指标和微平均 F1 指标。观察可知，PSO 在将数据投影到 1 维后再分类

的分类效果最优。

图 2 比较了采用多目标粒子群算法(MOPSO)进行降维后再分类的分类结果。其中，2-a 和 3-a 分别表示采用多目标优化投影指标(在非劣解集合中选取最优的 K 个投影方向)，将数据投影到 2 维和 3 维再分类的分类结果。观察可知，MOPSO 在 2-a 投影方式上取得了最好的分类结果，这说明了从非劣解集中选取的投影方向能够更好地表征原数据本身的结构。

图 3 比较了采用混沌粒子群算法优化多目标投影指标(MOCPSO)的分类结果。2-a 和 3-a 所表示的意思如前文所述。观察可知，MOCPSO 同样在 2-a 上取得了最优的实验结果。这一结果也证明了将数据投影到多维(2 维)能提高文本分类的性能。

4.2 不同分类模型之间的比较

为了进一步证明本文提出模型的有效性，实验比较了不同的分类模型的分类结果的宏平均(Macro Average)F1 指标和微平均(Micro Average)F1 指标。表 1 给出了不同的分类模型的实验结果。其中，LIBSVM 表示直接用 LIBSVM 分类器对高维数据分类；GA、SA、MCPSO 分别表示使用遗传算法、模拟退火、MCPSO 优化投影指标后将数据投影到 1 维后，用 LIBSVM 分类；PSO_1 表示用 PSO 优化后将数据投影到 1 维后，用 LIBSVM 分类；MOPSO_2-a 表示采用多目标优化投影指标，使用 PSO 优化将数据投影到 2 维后再使用 LIBSVM 分类器进行分类；MOCPSO_2-a 表示采用多目标优化投影指标，使用 CPSO 优化将数据投影到 2 维后再使用将数据投影到 2 维后再用 LIBSVM 分类器进行分类。

表 1 不同模型的分类结果比较

分类模型	MacroF1	MicroF1
LIBSVM	0.456	0.593
GA	0.550	0.610
SA	0.470	0.601
MCPSO	0.610	0.635
PSO_1	0.593	0.597
MOPSO_2-a	0.628	0.650
MOCPSO_2-a	0.693	0.701

由表 1 可知，MOCPSO_2-a 取得了最好的分类结果，这也证明了本文提出的模型的有效性；MOCPSO_2-a 和 MOPSO_2-a 的分类结果都优于 PSO_1，这说明采用多目标投影指标将数据投影到多维能在一定程度上提高文本分类的性能；直接用 LIBSVM 分类器对高维数据分类的分类结果是最差的，这说明了降低维度能有效地提高分类性能。当然，这个结论只是基于本文实验采用的 5 个新闻组数据。其结果还需要在其他数据集上进一步验证。

5 结束语

本文提出了一种基于 CPSO 的多目标投影寻踪文本分类模型。相对于传统的投影寻踪文本分类模型，本文做了以下 3 个方面的改进：(1)提出了一种多目标的投影指标，通过优化该投影指标可以更容易得到令人感兴趣的低维数据结构；(2)考虑了将数据投影到多维对文本分类的影响；(3)针对多目标投影指标，用混沌粒子群算法对投影指标最优化。后续研究将对多目标优化算法进行改进，以进一步提高分类性能。

参考文献

[1] Sebastiani F. Machine Learning in Automated Text Categorization[J]. ACM Computing Surveys, 2002, 34(1): 1-55.

[2] Lee E K, Cook D, Klinke S, et al. Projection Pursuit for Exploratory Supervised Classification[EB/OL]. [2012-05-01]. http://edoc.hu-berlin.de/series/sfb-649-papers/2005-26/PDF/26.pdf.

[3] Posse C. Projection Pursuit Discriminant Analysis for Two Groups[J]. Communications in Statistics——Theory and Method, 1992, 21(1): 1-19.

[4] Wei Huang, Zhang Xingnan. Projection Pursuit Flood Disaster Classification Assessment Method Based on Multi-Swarm Cooperative Particle Swarm Optimization[J]. Journal of Water Resource and Protection, 2011, 3(6): 415-420.

[5] 万中英, 王明文, 廖海波. 基于投影寻踪的中文网页分类算法[J]. 中文信息学报, 2005, 19(4): 60-67.

[6] 万中英, 廖海波, 王明文. 遗传-粒子群投影寻踪模型[J]. 计算机工程与应用, 2010, 46(20): 210-240.

[7] 廖海波, 万中英, 王明文. 免疫进化的投影寻踪模型在文本分类中的应用[J]. 广西师范大学学报: 自然科学版, 2011, 29(1): 123-128.

[8] Gholamian M R, Ghomi S M T F, Ghazanfari M. A Hybrid System for Multi-objective Problems——A Case Study in NP-hard Problems[J]. Knowledge-Based Systems, 2007, 20(4): 426-436.

[9] Kennedy J, Eberhart R. Particle Swarm Optimization[C]//Proc. of International Conference on Evolutionary Computation. Anchorage, USA: IEEE Press, 1995: 15-19.

[10] Zaslavskii G M. The Simplest Case of a Strange Attractor[J]. Physical Letters A, 1978, 69(3): 145-147.

[11] Alatas B, Akin E, Ozer A B. Chaos Embedded Particle Swarm Optimization Algorithms[J]. Chaos, Solitons & Fractals, 2009, 40(4): 1715-1734.

[12] Zhao Xianzhang, Zeng Junfang, Gao Yibo, et al. Particle Swarm Algorithm for Classification Rules Generation[C]// Proc. of ISDA'06. Jinan, China: [s. n.], 2006: 957-962.

编辑 金胡考