

面向大数据的在线特征提取研究

许烁娜 曾碧卿 熊芳敏

(华南师范大学软件学院 佛山 528225)

摘 要 在大数据环境下,当利用机器学习算法对训练样本进行分类时,训练数据的高维度严重制约了分类算法的性能。文中应用 L1 准则的稀疏性,提出了一种在线特征提取算法,并用该算法对训练实例进行分类。利用公开数据集对算法的性能进行了分析,结果表明,提出的在线特征提取算法能准确地对训练实例进行分类,因而能更好地适用于大数据环境下的数据挖掘。

关键词 大数据,机器学习,在线特征提取,算法

中图法分类号 TP312 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.09.045

Big Data Oriented Online Feature Extraction

XU Shuo-na ZENG Bi-qing XIONG Fang-min

(School of Software, South China Normal University, Foshan 528225, China)

Abstract In big data, the high dimension of training samples makes it difficult for classifying these samples during data mining. Applying the sparsity of the L1 norm, this paper proposed an online feature selection algorithm, and used this algorithm to classify the training samples. Experiments on public datasets show that the proposed online feature selection algorithm has better prediction accuracy than related work, and thus can be applicable to data mining for big data.

Keywords Big data, Machine learning, Online feature extraction, Algorithm

1 引言

特征提取是数据挖掘和机器学习领域的重要研究内容之一。特征提取的目标是从相关特征集中提取一个子集来建立一个有效的预测模型。通过移除不相关的和冗余的特征,特征提取可以缓解维度灾难、提高泛化性能、加速学习过程、提高模型的交互性,从而提高预测模型的性能。特征提取已广泛应用于各种应用领域,尤其用来解决高维数据面临的问题。

在现有的特征提取研究中,大多数研究假设训练实例的所有特征都是已知的,并且整个特征提取的过程是以离线的学习方式进行的。然而这种假设并不一直成立,在现实世界应用中,训练实例可能以在线的方式连续到达,或者收集训练数据的所有信息代价非常高。例如,在一个在线垃圾邮件检测系统中,训练数据经常顺序到达,因此部署一个及时、有效并且可扩展的批量特征提取方法是不可能的。在生物信息学中,获取每个训练实例的所有属性信息的代价非常高,因此离线的学习方式也是不可取的。

本文研究了在线特征提取,即通过在线学习的方法提取训练数据的特征,其目标是设计一个只含有数量很少且固定的特征集的在线分类器。在现实应用中,为了对大数据进行挖掘,必须要高效地识别数据集的相关特征集,然后通过在线学习过程建立准确的预测模型。本文提出了一种有效的在线

特征提取算法,分析了算法的特性,并通过大量的实验验证了该算法的性能。

2 相关工作

在机器学习和数据挖掘领域,在线特征提取包括在线学习过程和特征提取两个部分。本文对在线学习和特征提取的相关研究进行了介绍。

在在线学习算法中,感知器算法^[1,2]是使用最广泛的算法之一。感知器算法首先建立一个线性模型,然后不断地输入训练样本。当输入的训练样本没有正确分类时,感知器模型通过调整参数,使其对现有的样本全部正确分类。除了感知器算法以外,研究人员提出了大量的在线学习算法^[3-7],这些算法通常满足最大边际学习原则。Passive-Aggressive (PA)^[8]算法指出,在新实例到达的时候,分类器的更新应该满足前后两次的差值最小化。PA算法的局限在于其仅仅适用于一阶信息。为了使PA算法适用于二阶信息,文献[8-10]提出了置信加权在线学习算法。上述在线学习算法在分类时考虑了数据的全部特征,因而只能处理数据量很小的数据集。本文研究的是大数据环境下的在线学习,即每个数据包含大量的特征,因此必须对特征集合进行选取,提取那些最具代表性的特征集合。

特征提取是从数据的整个特征空间中提取出最相关的特

到稿日期:2013-11-18 返修日期:2014-03-26 本文受国家自然科学基金(71272144),广州市科技计划项目(2013KP084),广东省自然科学基金项目(8151063101000040)资助。

许烁娜(1974—),女,硕士,讲师,主要研究方向为大数据、社会网络分析,E-mail: xsn_nhxy@163.com;曾碧卿(1969—),男,博士,教授,主要研究方向为网络计算、分布式处理、并行 I/O、云计算;熊芳敏(1979—),硕士,讲师,主要研究方向为人工智能、蚁群算法、粒子群算法。

征子集,从而在保证预测模型具有一定准确性的前期下尽量提高模型的预测性能。特征提取的方法有很多,根据提取方法的不同,可以分为3类:过滤方法、包装方法和嵌入式方法。过滤方法^[11-13]依赖于数据的特征,如相关、距离和信息,而不包含任何的学习方法。包装方法^[14]需要一个预先定义好的学习算法来评估提取的特征集的性能。通俗地说,包装方法搜索整个特征空间,然后选择适合于预先定义好的学习算法的特征集合。包装方法虽然能提高预测模型的性能,但是计算复杂度却很高。在嵌入式方法^[15-17]中,特征选择过程和模型的训练过程同时进行。嵌入式方法的计算复杂度明显低于包装方法,并能为学习算法提供适合的特征子集,但是提取出的特征子集并不一定适合其它的学习算法。

文献[18-20]进行了在线特征提取的研究。这些研究假设所有的特征一次性到达,在学习过程开始前训练样本是可用的,其目标是在观测到的特征集中选取一个特征子集,并返回适合的模型。而本文关注的是训练样本顺序到达时的在线学习过程。文献[21,22]研究了稀疏的在线学习过程,它们的目标是通过学习建立一个稀疏的线性分类器。稀疏的线性分类器要求分类器是稀疏的,与之不同的是,本文严格限制了分类器中非0元素的个数。

3 在线特征提取算法

3.1 问题定义

本文研究的是二分类器下的在线特征提取。训练样本集 $D_t = \{(x_i, y_i) | i=1, \dots, T\}$ 中的元素以顺序的形式依次进行分类,其中 $x_i \in \mathbb{R}^d$ 是一个 d 维的向量, $y_i \in \{+1, -1\}$ 。 d 是一个非常大的整数,为了提高计算效率,需要为分类器提供一个相对小的特征子集。

该问题的形式化表示为,在每一个测试时间 t ,分类器 $w_t \in \mathbb{R}^d$ 通过线性函数 $\text{sgn}(w_t^T x_t)$ 对测试实例 x_t 进行分类。为了提高算法的性能,要求分类器 w_t 包含不超过 B 个的非0元素,即

$$\|w_t\|_0 \leq B \quad (1)$$

其中, B 是预先定义好的正整数。在分类过程中,测试实例 x_t 的特征最多被用到 B 个。我们称该问题是在线特征提取问题,目标是设计一个有效的在线特征提取策略,使得分类的误差最小。为了便于讨论,我们假设 $\|w_t\|_2 \leq 1, t=1, \dots, T$ 。

一种最直接的在线特征提取算法是通过感知器算法进行修改。在感知器算法中,在进行第 t 个测试实例时,可以对分类器 w_t 进行截断,即保留 w_t 中绝对值最大的 B 的元素,其它的元素设为0。然后应用截断的分类器进行后续的计算。虽然应用截断的感知器算法可以进行在线特征提取,但是它并不能保证被截断的元素对应的测试实例的对应属性值足够小,因此会带来很大的分类误差。

3.2 基于稀疏映射的在线特征提取算法

本文应用 L_1 准则的稀疏性原理,提出了基于稀疏映射的在线特征提取算法。 L_1 准则的命题如下:

命题 1^[23] 对于 $q > 1, x \in \mathbb{R}_d$, 则如下不等式成立:

$$\|x - x^m\|_q \leq \xi_q \|x\|_1 (m+1)^{1/q-1}, m=1, \dots, d \quad (2)$$

其中, ξ_q 是一个仅与 q 相关的常数, x^m 表示仅含有 x 中 m 个最大值的向量,其它的值设为0。

该命题的物理意义是,当一个向量 x 在一个 L_1 球中时, x 中的大多数的数值都聚集在它的那些最大元素的周围,因此,通过 L_q 准则将那些小的元素移除对原始向量的改变很小。我们将 L_1 球约束加入到在线特征提取分类器中,以满足

$$\Delta_R = \{w \in \mathbb{R}^d : \|w\|_1 \leq R\} \quad (3)$$

基于上述想法,本文设计了一种基于稀疏映射的在线特征提取算法。在线学习器维持一个线性分类器 W_t, W_t 中包含不超过 B 个的非0元素。当训练实例 (x_t, y_t) 被错误分类后,分类器通过梯度下降法进行更新,然后进行映射。映射完成后,对分类器的约束规则进行更新。如果分类器 $\hat{W}_{t+1}$ 有超过 B 个的非0元素,只保留其中绝对值最大的 B 个非0元素,其它的值设为0。算法细节见算法1和算法2。

算法1 在线特征提取(OFS)

输入: λ : 规范化参数; η : 步长; B : 特征提取数目

1. $W_t = 0$;
2. For $t=1$ to T do
3. 接收训练实例 x_t ;
4. 预测分类 $\text{sgn}(W_t^T x_t)$;
5. 得到分类结果 y_t ;
6. If $y_t W_t^T x_t \leq 1$ then
7. $\hat{W}_{t+1} = (1 - \lambda\eta) W_t + \eta y_t x_t$;
8. $\hat{W}_{t+1} = \min\{1, \frac{1}{\sqrt{\lambda}}\} \frac{\hat{W}_{t+1}}{\|\hat{W}_{t+1}\|_2}$;
9. $W_{t+1} = \text{Truncate}(\hat{W}_{t+1}, B)$;
10. Else
11. $W_{t+1} = (1 - \lambda\eta) W_t$;
12. End if;
13. End for;

在算法1中,我们首先初始化线性分类器 W_t 为0,然后对每个训练实例 x_t 应用符号函数 $\text{sgn}(W_t^T x_t)$:

$$\text{sgn}(s) = \begin{cases} 1, & s \geq 0 \\ -1, & s < 0 \end{cases} \quad (4)$$

对结果进行分类预测,并得到分类结果 y_t 。最后,根据预测结果对线性分类器进行改进。当 $y_t W_t^T x_t > 1$ 时,我们减小分类器的线性参数,使得 $W_{t+1} = (1 - \lambda\eta) W_t$; 当 $y_t W_t^T x_t \leq 1$ 时,增大分类器的线性参数,使得 $\hat{W}_{t+1} = (1 - \lambda\eta) W_t + \eta y_t x_t$,

并找到满足 $\min\{1, \frac{1}{\sqrt{\lambda}}\} \frac{\hat{W}_{t+1}}{\|\hat{W}_{t+1}\|_2}$ 的最小值。最后将最小

值 $\hat{W}_{t+1}$ 进行截断,使之保留绝对值大的前 B 个元素保留,截断函数 $\text{Truncate}(\hat{W}, B)$ 见算法2。

算法2 $W = \text{Truncate}(\hat{W}, B)$

1. If $\|\hat{W}\|_0 > B$ then
2. //将 $\hat{W}^B$ 中绝对值大的前 B 个元素保留,其它设为0
3. $W = \hat{W}^B$;
4. Else
5. $W = \hat{W}$;
6. End if;

4 实验

在本节,我们对本文提出的算法性能进行测试。我们对本文提出的在线特征提取(OFS)与基于截断感知器算法的分类器(PE_{trun})、随机特征提取(RAND,在所有的特征集合中随机选取 B 个特征)进行了对比。

4.1 数据集

本文使用的数据集全部为公开数据集,这些数据集可以在 LIBSVM 网站¹⁾和 UCI²⁾机器学习知识库中下载。表 1 列出了本文使用的 3 个数据集的详细信息。

表 1 数据集基本信息表		
数据集	个数	维度
magic04	19020	10
german	1000	24
spambase	4601	57

4.2 实验环境与参数选择

本实验在一台 4 核 4G 内存的个人 PC 上运行,算法实现语言为 Java,操作系统为 Ubuntu10。对每一个数据集,我们将提取 10%的特征,即 $0.1 * \text{维度}$,规范化参数 $\lambda=0.01$,学习效率参数 $\eta=0.2$,并且所有的参数选取方法相同。对每个数据集,我们随机进行 10 次实验,并用平均值来代表该数据集的最终结果。

4.3 实验结果

首先,实验对比了 3 种算法的在线预测性能。在 3 个数据集中,将维度的 10%作为特征提取出来,然后将该分类结果与全部属性的分类结果进行比较,然后对 10 次实验的错误次数进行平均,结果见表 2。

表 2 平均预测错误数对比

算法	german	magic04	spambase
RAND	472±11	8689±58	1827±45
PE_{trun}	489±29	8153±79	1294±66
OFS	432±13	6023±1342	913±157

从表 2 可以看出,在 3 种算法中,RAND 算法在 3 种数据集上的错误率都最高,这说明进行在线特征提取是必要的。在另外两种算法中,本文提出的 OFS 算法的错误率小于 PE_{trun} ,因此能很好地进行学习算法的在线特征提取。

为了进一步验证在线特征提取的性能,我们观察了 3 种算法在不同的数据集中的平均错误率随着训练样本数目的变化,实验结果见图 1—图 3。从这些图中可以看到,随着样本数目的增加,预测的平均错误率下降;在 3 个数据集中,OFS 算法的平均错误率都低于其它两种算法。

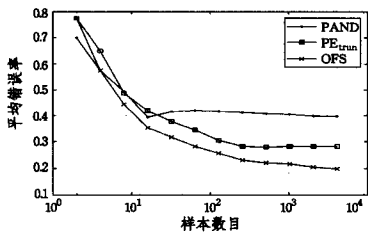


图 1 算法性能评估(spambase)

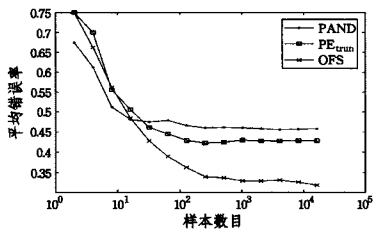


图 2 算法性能评估(magic04)

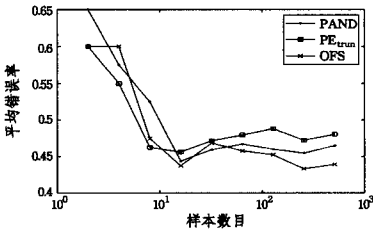


图 3 算法性能评估(german)

最后,我们观察了算法在不同的特征提取比例下的分类准确性,实验结果见图 4—图 6。当我们增加训练样本的特征提取比例时,算法的分类准确性也随着提高,并且本文提出的 OFS 算法在 3 个数据集下的预测准确性都明显优于其它两种算法,因此更适用于在线特征的提取。

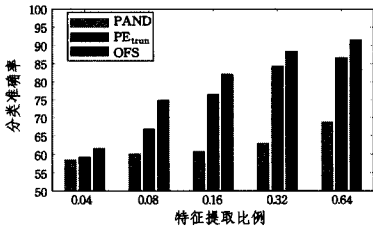


图 4 分类准确性评估(spambase)

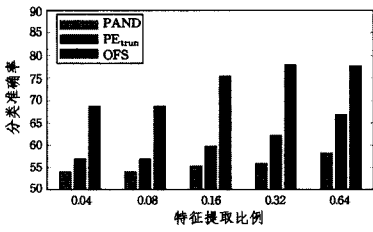


图 5 分类准确性评估(magic04)

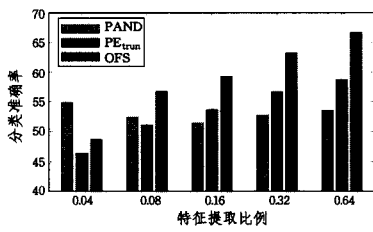


图 6 分类准确性评估(german)

结束语 在机器学习过程中,需要对训练数据进行分类,然而在大数据环境下,数据的维度非常高。为了提高学习算法的分类效率,需要对训练数据的特征进行提取,以选择少量能代表该数据的特征子集。本文应用 L1 准则的稀疏性,提出了一种在线特征提取算法,并应用该算法对训练实例进行

¹⁾ <http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/>
²⁾ <http://www.ics.uci.edu/~mllearn/MLRepository.html>

分类。实验结果表明,本文提出的在线特征提取算法能准确地对训练实例进行分类,因此能更好地适用于大数据环境下的数据挖掘应用。

参考文献

- [1] Sivaram G, Hermansky H. Sparse multilayer perceptron for phoneme recognition[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 23-29
- [2] Ahmad M Y, Mohamed A, Yusof M, et al. Colorectal cancer image classification using image pre-processing and multilayer Perceptron[C]//2012 International Conference on Computer & Information Science (ICIS). IEEE, 2012, 1: 275-280
- [3] Filipovych R, Resnick S M, Davatzikos C. JointMMCC: Joint Maximum-Margin Classification and Clustering of Imaging Data[J]. IEEE Transactions on Medical Imaging, 2012, 31(5): 1124-1140
- [4] Singh A, Ahuja N, Moulin P. Online learning with kernels: Overcoming the growing sum problem[C]//2012 IEEE International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2012, 1-6
- [5] Ralaivola L. Confusion-Based Online Learning and a Passive-Aggressive Scheme[C]//Advances in Neural Information Processing Systems 25. 2012, 3293-3301
- [6] Zhao P, Hoi S C, Jin R. Double updating online learning [J]. Journal of Machine Learning Research, 2011(12): 1587-1615
- [7] Wang J, Zhao P, Hoi S C. Exact soft confidence-weighted learning[C]//ICML. 2012
- [8] Crammer K, Dredze M, Pereira F. Confidence-weighted linear classification for text categorization[J]. The Journal of Machine Learning Research, 2012, 98888: 1891-1926
- [9] Wang J, Zhao P, Hoi S C H. Exact soft confidence-weighted learning[C]//Proceedings of the 29th International Conference on Machine Learning (ICML-12). 2012, 121-128
- [10] 王维, 张鹏涛, 谭营, 等. 一种基于人工免疫和代码相关性的计算机病毒特征提取方法[J]. 计算机学报, 2011, 34(2): 204-215
- [11] Ditzler G, Rosen G, Polikar R. Information theoretic feature selection for high dimensional metagenomic data[C]//2012 IEEE International Workshop on Genomic Signal Processing and Statistics, (GENSIPS). IEEE, 2012, 143-146

- [12] 王蕊, 冯登国, 杨铁, 等. 基于语义的恶意代码行为特征提取及检测方法[J]. 软件学报, 2012, 23(2): 378-393
- [13] Wu Y C, Yang J C. A Weighted Cluster-based Chinese Text Categorization Approach: Incorporating with Word Clusters[C]//2012 IIAI International Conference on Advanced Applied Informatics (IIAIAI). IEEE, 2012, 279-282
- [14] Wang G, Song Q, Sun H, et al. A Feature Subset Selection Algorithm Automatic Recommendation Method[J]. J. Artif. Intell. Res. (JAIR), 2013, 47: 1-34
- [15] Snelson E, Ghahramani Z. Variable noise and dimensionality reduction for sparse Gaussian processes[C]//Proceedings of the Twenty-Second Conference Annual Conference on Uncertainty in Artificial Intelligence. 2006, 461-468
- [16] Chandrasekaran V, Jordan M I. Computational and statistical tradeoffs via convex relaxation[C]//Proceedings of the National Academy of Sciences. 2013
- [17] 业巧林, 赵春霞, 陈小波. 基于正则化技术的对支持向量机特征选择算法[J]. 计算机研究与发展, 2011, 48(6): 1029-1037
- [18] Bibicu D, Moraru L, Biswas A. Thyroid nodule recognition based on feature selection and pixel classification methods[J]. Journal of Digital Imaging, 2013, 26(1): 119-128
- [19] Yu K, Ding W, Simovici D A, et al. Mining emerging patterns by streaming feature selection[C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2012, 60-68
- [20] Langford J, Li L, Zhang T. Sparse online learning via truncated gradient[J]. The Journal of Machine Learning Research, 2009 (10): 777-801
- [21] Combettes P L, Vũ B C. Variable metric forward-backward splitting with applications to monotone inclusions in duality[J]. Optimization, 2014, 63(9): 1289-1318
- [22] Binev P, Dahmen W, DeVore R, et al. Compressed sensing and electron microscopy[M]//Modeling Nanoscale Imaging in Electron Microscopy. Springer US, 2012, 73-126
- [23] Donoho, Leigh D. Compressed sensing[J]. IEEE Transactions on Information Theory, 2006, 52(4): 1289-1306
- [24] 林颖. 闭合序列模式的一种增量挖掘算法[J]. 重庆理工大学学报: 自然科学版, 2011, 25(6): 95-100

(上接第 231 页)

研究的热点。传统的博弈理论假定所有参与人关于博弈的信息都是完全的、精确的。然而,这一严格的假定在真实的博弈环境中有时候会得不到满足。在一个现实的博弈中,由于信息的不完整和知识的缺乏,某个参与人可能并不确切地知道对应于某个战略组合的支付值。传统的博弈理论不能解决这种博弈中存在的不确定性的问题。本文对支付值为格值的二人零和矩阵博弈进行研究。基于该类型博弈的特殊性,定义了新的均衡解,研究了均衡解的性质,给出了获得均衡解的方法,获得了均衡解存在的充分必要条件,为解决支付值存在着不可比较性的矩阵博弈提供了一种新的方法。

参考文献

- [1] Von N J, Morgenstern O. Theory of Games and Economic Behavior[M]. Princeton: Princeton University Press, 1944
- [2] Campos L, Gonzalez A. Fuzzy matrix games considering the criteria of the players[J]. Kybernetes, 1999, 20: 275-289
- [3] Yager R R. A Procedure for ordering fuzzy numbers in the unit

interval[J]. Information Sciences, 1981, 24: 143-161

- [4] Larbani M. Non cooperative fuzzy games in normal form: A survey[J]. Fuzzy Sets and Systems, 2009, 160: 3184-3210
- [5] Blair C, Roth A E. An extension and simple proof of a constrained lattice fixed point theorem [J]. Algebra Universalis, 1979, 9: 131-132
- [6] Roth A E. A lattice fixed-point theorem with constraints [J]. Bulletin of the American Mathematical Society, 1975, 81: 136-138
- [7] Tarski A. A lattice-theoretical fix-point theorem and its applications[J]. Pacific Journal of Mathematics, 1955(5): 285-309
- [8] Zimper A. A fixed point characterization of the dominance-solvability of lattice games with strategic substitutes[J]. International Journal of Game Theory, 2007, 36: 107-117
- [9] 杨丽, 徐扬. 基于矩阵蕴涵运算的格值模糊概念格构造方法[J]. 计算机科学, 2009, 36: 264-267
- [10] Xu Y, Liu J, Zhong X M, et al. Lattice-valued matrix game with mixed strategies for intelligent decision support[J]. Knowledge-Based Systems, 2012, 32: 56-64