

基于加权 Bayes 分类器的流数据在线分类算法研究

卢惠林

(哈尔滨工业大学计算机学院 哈尔滨 150001) (江苏省无线传感系统应用技术研发中心 无锡 214153)

摘要 传统的分类算法在对模型进行训练之前,需要得到整个训练数据集。然而在大数据环境下,数据以数据流的形式源源不断地流向系统,因此不可能预先获得整个训练数据集。研究了大数据环境下含有噪音的流数据的在线分类问题。将流数据的在线分类描述成一个优化问题,提出了一种加权的 Naïve Bayes 分类器和一种误差敏感的(Error Adaptive)分类器,并通过真实的数据集对提出的算法进行了验证。实验结果表明,文中提出的误差敏感的分类器算法在系统没有噪音的情况下分类预测的准确性要优于相关的算法;此外,当流数据中含有噪音时,误差敏感的分类器算法对噪音不敏感,仍然具有很好的预测准确性,因此可以应用于大数据环境下流数据的在线分类预测。

关键词 大数据,决策树,分类算法,流数据

中图法分类号 TP311 **文献标识码** A

Weighted Bayes Based Data Streaming Online Classification Algorithm

LU Hui-lin

(School of Computer, Harbin Industrial University, Harbin 150001, China)

(Jiangsu Engineering Technology Development Center on Wireless Sensor System, Wuxi 214153, China)

Abstract Traditional classification algorithms need to obtain the whole training dataset before training the model. However, for big data, data are streaming into the system sequentially, so it is impossible to obtain the whole training dataset beforehand. This paper studied the online classification problem in data streaming for big data. It first described the online classification problem as an optimization problem, then proposed a Weighted Naïve Bayes classifier and an Error Adaptive classifier, and at last, validated the efficiency of the proposed algorithm according to two real datasets. The experiments show that the prediction accuracy of our proposed algorithm is higher than related researches in non-noisy data streaming, and moreover, while data streaming is noisy, our algorithm still has better prediction accuracy, so it can be used in real online classification application in data streaming.

Keywords Big data, Decision tree, Classification algorithm, Data streaming

近些年,由于数据规模呈爆炸式增长,因此大数据成为研究人员广泛关注的话题。在数据流环境下,由于数据以流的方式不断到达系统,因此预处理和后处理的数据处理技术都不能有效应对。决策树是一种广泛使用的分类方法,其树形结构能很好地表示推理规则,并且解释性很好。在大数据挖掘中,由于数据量大,在应用决策树设计机器学习算法时会导致树的规模很大,为了压缩决策树的规模,并保持分类的准确性,必须设计新的决策树技术。

快速决策树(Very Fast Decision Tree, VFDT)系统是数据流挖掘设计的重要算法之一。在数据流应用中,快速决策树是轻量级的数据挖掘算法,当系统收集到足够的统计信息后,它可以在系统中从头开始实时构建决策树^[1]。在扫描数据时,快速决策树通过增量更新决策树来进行学习。快速决策树不同于传统的基于归纳的决策树,后者需要知道数据集的全部信息,而前者通过流的形式只需要部分信息,并且能增量更新^[2]。快速决策树的最大优点是可以进行数据流的实时挖掘。在传统的基于归纳的决策树中,系统需要存储所有

相关数据,当数据发生改变时,系统必须对决策模型进行重新训练。在数据流环境下,数据是没有边际的,传统的归纳决策树并不适用于这种情况,而快速决策树却可以根据数据流动态地更新决策树^[3]。在快速决策树算法中,最重要的挑战是数据流的数据质量问题,例如,数据网络发生故障会使数据流中的数据含有噪音,这将导致学习算法的不准确。

大数据环境下,数据以数据流的形式连续到达系统,本文研究了如何应用分类算法对流数据进行在线分类学习。

1 相关工作

分类算法是数据挖掘领域的重要研究内容,已有大量的相关研究。目前,使用最广泛的分类算法有基于决策树的分类算法、基于关联规则的分类算法^[4-6]、贝叶斯分类算法^[7-9]、神经网络分类算法^[10,11]等。本文采用基于决策树的分类算法,以下对基于决策树的分类算法及其在数据流上的相关研究做一介绍。

决策树通过树形规则对多维数据进行分类。每一个数据

项从树的根节点输入,然后根据一系列判断规则对数据项进行细分,直到其对应的节点为叶节点(即分类)。为了提高分类的准确性,Pfahringner 等人^[12]提出了一种同时优化多个树枝的分类算法。该算法利用那些优化的规则取代准确性低的规则,因此预测结果的准确性显著提高,然而在构建优化的树分支时增加了大量的计算,因此计算效率非常低。在决策树中,不活跃的分支仍然消耗计算资源,为此 Gama 等人^[13]提出了功能树叶的思想,即将部分功能移到树的叶节点。通过将部分功能移到叶节点,既可以提高系统的效率,又可以提高分类的准确性。该分类器在数据流环境下既能处理连续变量,也能处理离散变量,但是当数据流中的信息含有噪音或者偏好时,该算法的效果不佳。为了提高系统应对噪音的能力,Hashemi 等人^[14]提出了 Sigmoid 函数。在含有噪音和缺失数据的数据项中,该 Sigmoid 函数能准确地对数据项进行预测,但是却带来了额外的开销,导致算法的性能低。

在数据流挖掘中,决策树算法随着数据项的到来增量地构建决策树模型。在这种不稳定的环境下,计算资源显得相对有限。Yang 等人^[2]设计了一个基于 C 语言的挖掘高速改变的数据流工具箱——VFML。Gama 等人^[15]设计了一个基于 Java 语言的大数据分析软件——MOA。MOA 是一个非常著名的基于 WEKA 的开源数据挖掘软件。VFML 和 MOA 这两个平台在进行数据挖掘时都需要对 VFDT 的参数进行预先设定。然而在实际情况下,不同的预测模型需要的参数是不同的,很难进行预先定义。

MOA 是一个用户界面友好的开源平台,并提供了一些评价算法性能的方法。在 MOA 的基础上,研究人员对其进行了扩展,典型的扩展算法有 HOT、ASHT、ADAHOT 和 ADWIM。HOT^[16]是一个带有集合 VFDT 分类器的在线包装方法。在 VFDT 的基础上,ASHT^[17]算法增加了分割节点的最大值。当一个节点被分割后,如果被分割的节点超过了最大值,算法通过删除多余的节点来减小节点的规模。ADAHOT^[12]将投票机制与 HOT 算法相结合。在 ADAHOT 中,每个叶节点存储当前误差的估计值;在投票中,节点的权重与其误差成反比。此外,在假设滑动窗口的平均值不变的前提下,Bifet 和 Gavalda^[18]提出的 ADWIM 算法可以通过一个变长的滑动窗口检测近期内的数据变化。

2 基于决策树的增量分类算法

2.1 问题定义

在大数据中,每个数据项都有一个时间戳 t ,数据项 D_t 还有一个属性向量 X^t 和一个类别 y^t ,即

$$D_t = [X^t, y^t] \quad (1)$$

在一个有限的时间内到达的数据项的集合为 D_T ,其中 T 是 D_T 中所含有数据项的个数,且 $T > 0$ 。一个完整的数据流定义如下:

$$D_T = \sum_{t=1}^T D_t = \begin{bmatrix} X^1 & y^1 \\ \vdots & \vdots \\ X^T & y^T \end{bmatrix} \quad (2)$$

令 $H(\cdot)$ 为分类算法的启发式函数,当数据集合 D_T 有限时,训练模型可以对所有训练实例进行学习,然后获取一个全局的训练模型,并根据该训练模型对后续数据进行预测,其目标是使得预测结果的误差达到最小。该优化模型可定义如下:

$$\begin{cases} TR_{Global} = Train(D_T, H(\cdot)) \\ \hat{y}_k = Test(TR_{Global}, X^t) \\ Error_k = \begin{cases} 1, & \hat{y}_k \neq y_k^t \\ 0, & \hat{y}_k = y_k^t \end{cases} \\ Minimize \sum_{t=1}^T \sum_{k=1}^K Error_k^t \end{cases} \quad (3)$$

其中, $\hat{y}_k$ 表示应用模型 TR_{Global} 对 X^t 的分类预测值, k 为其所属的分类。

在大数据环境下,数据以流的形式源源不断地进入系统,因此不可能在获得全局数据 D_T 后再进行模型的训练。为此,需要设计增量的训练模型。当每一个 D_{t-1} 到达时,对模型进行训练,并用训练的结果对下次到来的数据进行预测。

$$\begin{cases} TR_{Incr} = Train(D_{t-1}, H(\cdot)) \\ \hat{y}_k = Test(TR_{Incr}, X^t) \\ Error_k = \begin{cases} 1, & \hat{y}_k \neq y_k^t \\ 0, & \hat{y}_k = y_k^t \end{cases} \\ Minimize \sum_{t=1}^T \sum_{k=1}^K Error_k^t \end{cases} \quad (4)$$

本文的目标是设计一个增量的启发式函数 $H(\cdot)$,使得在模型训练和预测过程中,模型的预测误差之和达到最小。

2.2 增量的决策树算法

在大数据环境下,设计增量的启发式函数,不仅要使模型的预测误差之和达到最小,还要有高的学习速度。本文通过决策树模型对启发式函数进行建模。

当新的数据项到达时,该数据项被存储在根节点,通过当前的决策树进行判别,将其分类在相应的叶节点。叶节点通过嵌入式分类器提高数据项的分类准确性。本文采用 4 个不同的分类器: Majority Class, Naïve Bayes, Weighted Naïve Bayes 和 Error Adaptive。

给定一个数据项 $D_t = [X^t, y^t]$, 它的分类结果为 $y_k^t = y^t$ 。对于 D_t 的每个属性 X_i^t ($1 \leq i \leq m$), 我们用一个大的类分布表 (Observed Class Distribution, OCD) 来存储第 i 属性的统计分布。假设 J 为属性 i 观测到的不同取值的个数, K 为分类的个数,那么 OCD 为一个 $J \times K$ 的矩阵,其中元素 n_{ijk} 表示数据项的第 i 个属性的取值为 X_i^t 时,被分在第 k 类中的数据项的个数,于是有,

$$OCD_i = \begin{bmatrix} n_{i11} & \cdots & n_{i1J} \\ \cdots & \ddots & \cdots \\ n_{iK1} & \cdots & n_{iJK} \end{bmatrix} \quad (5)$$

当给定属性 i ($1 \leq i \leq m$) 的值为 x_{ij} (x_{ij} 属于 J 个取值中的一个,即 $1 \leq j \leq J$) 时,它被分在类别 k 里的概率为

$$OCD_{ijk} = \frac{n_{ijk}}{\sum_{k=1}^K n_{ijk}}, (1 \leq i \leq m, 1 \leq j \leq J) \quad (6)$$

Majority Class^[13] (F^{MC}) 分类器应用最大概率值对数据项的属性进行分类,即返回概率最大的分类:

$$\operatorname{argmax} k = \{OCD_{ij1}, \cdots, OCD_{ijk}, \cdots, OCD_{ijK}\} \quad (7)$$

其中 $n_{ijk} > n_{ijk'} (1 \leq k' \leq K \wedge k' \neq k)$ 。

Naïve Bayes^[13] (F^{NB}) 分类器应用贝叶斯模型计算分类的

概率并返回概率最大值所在的分类,贝叶斯计算公式如下:

$$p_{ijk} = \frac{p(x_{ij} | y_k) \cdot p(y_k)}{p(x_{ij})} \quad (8)$$

Naïve Bayes 分类器的分类结果为

$$\operatorname{argmax} k = \{p_{ij1}, \dots, p_{ijk}, \dots, p_{ijK}\} \quad (9)$$

在 Naïve Bayes 分类器的基础上,本文提出加权的贝叶斯分类器(Weighted Naïve Bayes, F^{WNB}),其计算公式为

$$p_{ijk} = \text{OCD}_{ijk} \frac{p(x_{ij} | y_k) \cdot p(y_k)}{p(x_{ij})} \quad (10)$$

其分类结果仍然按照式(9)进行计算。

为了对分类器的预测结果进行调优,本文设计了一种误差敏感的(Error Adaptive, F^{EA})分类器,该分类器从上述 3 种分类器中选取一个误差最小的分类器。

在时刻 T ,系统共进入的数据项个数为 $T = \sum_{k=1}^K n_{ijk}$,分类器 F 对分类 y_k 的计算误差为

$$\text{ERR}(F, y_k) = \sum_{t=1}^T \text{Error}_t^k \quad (11)$$

Error Adaptive 分类器通过上述 3 种分类器的最小误差来对数据进行分类,

$$\operatorname{argmin} F^{EA} = \{\text{ERR}(F^{MC}, y_k), \text{ERR}(F^{NB}, y_k), \text{ERR}(F^{WNB}, y_k)\} \quad (12)$$

3 实验

3.1 评价标准

在对算法的性能进行评价时,我们利用了算法的分类准确性,即正确的分类数目占所有预测数据的比例:

$$\text{Precision} = \frac{\sum_{t=1}^T \text{Error}_t^k}{T} \quad (13)$$

3.2 数据集

在数据集的选择过程中,我们从 UCI 的机器学习知识库中下载了两个数据集 Person Activity 和 Cover Type。数据集的基本信息见表 1。

表 1 数据集基本信息表

数据集名称	类型	#Nom	#Num	#Cls	Size
Person Activity	Mix	2	3	22	164860
Cover Type	Mix	42	12	7	581012

Person Activity 于 2010 年发布,它记录了来自于人体的传感器发来的数据。该数据集的每个实例包括 2 个类别属性(传感器标识和时间)、3 个来自于传感器的数值属性。通过连续的数据实例,可以预测 22 个人类活动,如走路、弯腰、跑步等。Cover Type 数据集发布于 1999 年,其中每个数据实例包含 42 个类别属性、12 个连续属性。在 Cover Type 数据集中,数据共分为 7 个分类。

3.3 结果分析

首先,我们对比了 F^{MC} 、 F^{NB} 、 F^{WNB} 和 F^{EA} 4 种算法在不同数据集下的分类准确性,对比结果见图 1。从图 1 可以看出,在 Person Activity 数据集中, F^{WNB} 的准确性高于 F^{MC} 和 F^{NB} ;在 Cover Type 数据集中, F^{NB} 的准确性高于 F^{WNB} 和 F^{MC} ; F^{WNB} 在两个数据集下的准确性都高于 F^{NB} ; F^{EA} 根据误差选取最优的分类算法,因此其在两个数据集下的准确性都明显高于其它 3 种算法。

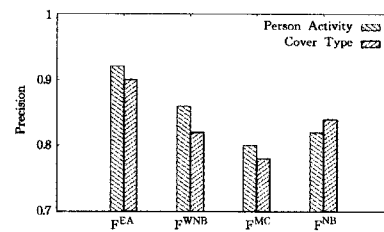


图 1 分类器准确性比较

其次,我们将本文提出的 Error Adaptive 分类算法 F^{EA} 与 HOT 算法、ASHT 算法、ADAHOT 算法和 ADWIM 算法进行了对比,结果见图 2。从图 2 可以看出,由于 F^{EA} 算法对分类结果进行了调优,因此其在两种数据集下的分类准确性都高于其他相关算法。

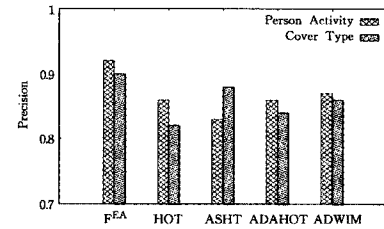


图 2 算法与相关研究的对比

最后,我们测试了上述 5 种算法对数据集中噪音的适应性。在实验所用的两个数据集中,分别以 10% 的概率注入随机噪音。在新数据即将到来时,我们以 10% 的概率随机向系统注入一个随机的数据项,该数据项的各项属性及所属分类分别在已知的属性集合和分类中随机选取。由于在数据流中注入了随机噪音,好的分类算法能有效地将其识别,并能正确分类。我们对注入了随机噪音的两个数据集的分类准确性进行了分析,结果见图 3。从图 3 可以看出,在数据集中含有噪音的情况下,本文提出的 F^{EA} 分类算法的分类预测准确性只有略微的下降,而其他算法的分类预测准确性明显下降。因此本文提出的 F^{EA} 分类算法在大数据环境下能更好地进行流数据的在线学习分类。

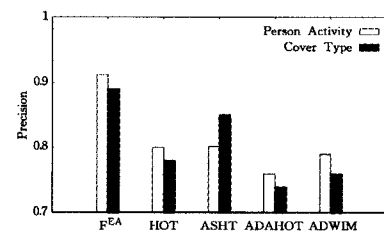


图 3 算法在含有噪音情况下的准确性对比

结束语 随着物联网技术的不断发展,大量的应用领域都采用传感器作为获取信息的重要手段。在传感器数据中,数据以数据流的形式源源不断地流向系统。传统的机器学习分类算法在假设预先得到整个数据集的前提下,进行数据的分类。这种假设在流数据环境下是不成立的,因为流数据的结束就意味着整个应用的终止,这时即使可以对数据进行预测,得到的结果也是滞后的、无用的。

本文研究了流数据环境下含有噪音的流数据的在线分类算法。首先将流数据的在线分类描述成一个优化问题,其次提出了一种加权的 Naïve Bayes 分类器和一种误差敏感的(Error Adaptive)分类器,最后通过真实数据集对算法进行了

(下转第 234 页)

- [9] Paul S. Andrews. An Investigation into Mutation Operators for Particle Swarm Optimization[C]//IEEE Congress on Evolutionary Computation. 2006:1044-1050
- [10] 吴祁宗. 运筹学与最优化方法[M]. 北京:机械工业出版社, 2003:114-117
- [11] 赵小平. 差商最速下降法及其收敛性[J]. 华东化工学院学报, 1992,18(6):807-809
- [12] Yao Xin, Liu Yong, Lin Guang-ming. Evolutionary Programming Made Faster[J]. IEEE Transactions on Evolutionary Computation, 1999,3(2):82-102
- [13] Shi Yu-hui, Eberhart R C. A Modified Particle Swarm Optimizer [C]//Proceeding of IEEE International Conference on Evolutionary Computation. 1998,69-73
- [14] Poli R, Kennedy J, Blackwell T. Particle swarm optimization[J]. Swarm Intelligence Journal, 2007,1:33-57

(上接第 226 页)

- [14] Nakano K, Nishimura S. Deriving Event-Based Document Transformers from Tree-Based Specifications [J]. Electronic Notes in Theoretical Computer Science, 2001,44:181-205
- [15] Becker O. Transforming XML on the Fly[C]//XML Europe. 2003
- [16] Jiang H, Ho H, Popa L, et al. Mapping-driven XML transformation[C]//WWW, Banff, Alberta, Canada, 2007:1063-1072
- [17] Frisch A, Nakano K. Streaming XML transformations using term rewriting[C]//PLAN-X, Nice, France, 2007:2-13
- [18] Dvoráková J, Zavoral F. A Low-Memory SSXT Algorithm for XSLT Transformations[J]. Journal of Information Assurance and Security, 2008,3:230-239
- [19] Tanaka Y, Ito K, Fujima J. Meme media for clipping and combining Web resources[C]//International Conference on World Wide Web. Hong Kong, 2001:201-210
- [20] Fagin R, Haas L M, Hernández M, et al. Clio: Schema Mapping Creation and Data Exchange[C]//Lecture Notes in Computer Science, vol. 5600. Heidelberg Berlin: Springer, 2009:198-236
- [21] 宋玲, 吕强, 邓薇, 等. 基于语义和结构的 XML 文档相似度的计算方法[J]. 中文信息学报, 2012,26(5):59-64
- [22] 陈谊, 侯望, 新吉乐, 等. 基于 XML 和关系数据库的可视化工作流系统[J]. 系统仿真学报, 2012,24(1):110-116

(上接第 229 页)

测试。实验结果表明,本文提出的误差敏感的分类器算法在系统没有噪音的情况下分类预测的准确性要优于相关的算法;此外,误差敏感的分类器算法对噪音不敏感,当流数据中含有噪音时仍然具有很好的预测准确性。因此,该算法可以应用于大数据环境下数据流的在线分类预测。

参考文献

- [1] Domingos P, Hulten G. Mining high-speed data streams [C]//Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery And Data Mining. ACM, 2000:71-80
- [2] Yang H, Fong S. Moderated VFDT in stream mining using adaptive tie threshold and incremental pruning[M]//Data Warehousing and Knowledge Discovery. Springer, 2011:471-483
- [3] Hulten G, Spencer L, Domingos P. Mining time-changing data streams[C]//Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery And Data Mining. 2001:97-106
- [4] Li W, Han J, Pei J. CMAR: Accurate and efficient classification based on multiple class-association rules[C]//IEEE International Conference on Data Mining. ACM, 2001:369-376
- [5] Han J. CPAR: Classification based on predictive association rules [OL]. <http://sci2s.ugr.es/keel/pdf/algorithm/congreso/2003-Yin-CPAR.pdf>, 2003
- [6] Thabtah F, Cowling P, Peng Y. MCAR: multi-class classification based on association rule[C]//The 3rd ACS/IEEE International Conference on Computer Systems and Applications. IEEE, 2005
- [7] 詹英, 吴春明, 王宝军. 一种与缓冲区紧耦合的环形循环滑动窗口的数据流抽取算法[J]. 电子学报, 2011,39(4):2262-2267
- [8] 崔贲勋, 李梁, 王柯柯, 等. 关联规则挖掘中 Apriori 算法的研究与改进[J]. 计算机应用, 2010,30(11):2952-2955
- [9] 詹英, 吴春明, 王宝军. 基于 RCSW 的数据流速度异常检测算法研究[J]. 电子学报, 2012,40(4):674-680
- [10] 吴枫, 仲妍, 吴泉源. 基于增量核主成分分析的数据流在线分类框架[J]. 自动化学报, 2010,36(4):534-542
- [11] Tang L, Tian L F, Steward B L. Classification of broadleaf and grass weeds using gabor wavelets and an artificial neural network[J]. Transactions of the Asae, 2003,46(4):1247
- [12] Pfahringer B, Holmes G, Kirkby R. New options for hoeffding trees[M]//AI 2007: Advances in Artificial Intelligence. Springer, 2007:90-99
- [13] Gama J A O, Rocha R, Medas P. Accurate decision trees for mining high-speed data streams[C]//Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2003:523-528
- [14] Hashemi S, Yang Y. Flexible decision tree for data stream classification in the presence of concept change, noise and missing values[J]. Data Mining and Knowledge Discovery, 2009,19:95-131
- [15] Bifet A, Holmes G, Kirkby R, et al. Moa: Massive online analysis [J]. The Journal of Machine Learning Research, 2010,99:1601-1604
- [16] Oza N C. Online bagging and boosting[C]//2005 IEEE International Conference on Systems, Man And Cybernetics. IEEE, 2005:2340-2345
- [17] Bifet A, Holmes G, Pfahringer B, et al. New ensemble methods for evolving data streams[C]//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery And Data Mining. 2009:139-148
- [18] Bifet A, Gavalda R. Learning from time-changing data with adaptive windowing[OL]. <http://www.lsi.upc.edu/~abifet/TimevaryingE.pdf>
- [19] 王柯柯, 崔贲勋, 倪伟, 等. 基于单元的快速的大数据集离群数据挖掘算法[J]. 重庆邮电大学学报: 自然科学版, 2010,22(5):673-677