一种挖掘概念漂移数据流的选择性集成算法

关菁华 刘大有
(吉林大学符号计算与知识工程教育部重点实验室 长春 130012)
(吉林大学计算机科学与技术学院 长春 130012)

摘 要 提出一种挖掘概念漂移数据流的选择性集成学习算法。该算法根据各基分类器在验证集上的输出结果向量方向与参考向量方向之间的偏离程度,选择参与集成的基分类器。分别在具有突发性和渐进性概念漂移的人造数据集SEA和Hyperplane上进行实验分析。实验结果表明,这种基分类器选择方法大幅度提高了集成算法在处理概念漂移数据流时的分类准确性。使用error-ambiguity分解对算法构建的naive Bayes集成在解决分类问题时的性能进行了分析。实验结果表明,算法成功的主要原因是它能显著降低平均泛化误差。

关键词 概念漂移,选择性集成,朴素贝叶斯,error-ambiguity分解
中图法分类号 TP181 **文献标识码** A

Selected Ensemble of Classifiers for Handling Concept-drifting Data Streams

GUAN Jing-hua LIU Da-you
(Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China)
(College of Computer Science and Technology, Jilin University, Changchun 130012, China)

Abstract In data streams concept is often not stable but change with time. We proposed a selective integration algorithm OSEN (Orientation based Selected ENsemble) for handling concept drift data streams. This algorithm selects a near optimal subset of base classifiers based on the output of each base classifier on validation dataset. Our experiments with synthetic data sets simulating abrupt (SEA) and gradual (Hyperplane) concept drifts demonstrate that selective integration of classifiers built over small time intervals or fixed-sized data blocks can be significantly better than majority voting and weighted voting, which are currently the most commonly used integration techniques for handling concept drift with ensembles. This paper also explained the working mechanism of OSEN from error-ambiguity decomposition. Based on experiments, OSEN improves the generalization ability through reducing the average generalization error of the base classifiers constituting the ensembles.

Keywords Concept drift, Selective ensemble, Naive bayes, Error-ambiguity decomposition

1 引言

众多应用领域的数据不断增加,其包含的模式会随时间和应用环境而变化,这种现象称为"概念漂移"。目前国内外在处理概念漂移方面已进行了大量研究,提出了多种模式学习方法。它们可归结成两类[1]:基于实例选择和基于集成学习的方法。基于实例选择的方法尽可能选择与当前概念相关的实例进行学习。该方法使用一个全局分类器,从而遗忘了所有的历史信息,因此不能很好地处理概念漂移问题。

基于集成的方法通过保留过去学习到的概念,既避免了灾难性遗忘又避免了因保存大量实例所需占用的计算资源。为了处理概念漂移问题,这类方法需根据原有概念与当前数据的一致性动态删除一些旧的分类器,生成新的分类器。这类方法主要包括Littlestone等人提出的Weighted Majority (WM)[2]算法、Freund等人提出的Hedge[3]算法、Street和Kim[4]提出的算法、Kolter等人提出AddExp(Additive expert)集成算法[5]、M_ID4[6]和Wang H.等人提出的用于处理概念漂移数据流的集成算法[7]等。已有的基于集成的方法不能及时丢弃无用分类器,造成错误概念的干扰,影响分类预测结果。

本文提出了应用于挖掘概念漂移数据流的选择性集成学习算法OSEN(Orientation Selected ENsemble)。算法根据各基分类器在验证数据集上的输出结果向量与参考向量之间的角度来选择参与集成的基分类器。实验研究表明,此选择方法能显著降低挖掘概念漂移数据流时的集成泛化误差。本文使用error-ambiguity分解对OSEN构建的naive Bayes集成在解决分类问题时的性能进行了分析。实验结果表明,与不进行基分类器选择的集成方法相比,OSEN能够显著降低平

到稿日期:2009-02-20 返修日期:2009-06-02 本文受国家自然科学基金重大项目(60496321),国家自然科学基金项目(60373098,60573073),国家高技术研究发展计划项目(20060110Z2037),吉林省科技发展计划重大项目(20020303),吉林省科技发展计划项目(20030523),欧盟项目TH/Asia Link/010(111084)资助。

关菁华(1979-),女,博士研究生,研究方向为集成学习、贝叶斯网,E-mail:gjh_jlu@hotmail.com;刘大有(1942-),男,教授,博士生导师,研究方向为知识工程与专家系统、分布式AI与多Agent系统、不确定性推理、空间推理与GIS应用等。

均泛化误差。

2 挖掘概念漂移数据流的选择性集成学习算法

为降低参与集成的基分类器数目，提高集成的泛化性能，大量的选择技术已用于解决集成问题。对于一个包括 n 个基分类器的集成进行选择，需要在 2^n-1 非空子集空间进行搜索，选择能最小化泛化误差的子集，这是个 NP 搜索问题。为了解决这个问题，提出了众多启发式搜索方法。Zhou 等人[8]提出了 GASEN 算法，其运用遗传算法确定基分类器的权重，权重反映了个体神经网络在集成中的重要性，根据该权重选择部分神经网络参与集成；Giacinto 等人[9]的聚类神经网络集成方法，根据个体神经网络的输出对个体神经网络进行聚类，从每类中各选出一个个体神经网络参与集成；Bakker 等人[10]的聚类神经网络集成方法，也根据个体神经网络的输出对个体神经网络进行聚类，最后根据聚类的结果重新构造一个个体神经网络参与集成。这几种方法都取得了较好的集成效果。但这几种方法都将选择性集成技术应用于静态数据，而非动态流数据。

本文针对概念漂移数据流的特点，提出了 OSEN 算法。算法描述如图 1 所示。当一个新数据块到来时，算法根据新数据块构建一个分类器；将新数据块作为验证集，计算过去产生的分类器在验证集上的泛化误差；为避免过适应，新分类器的泛化误差通过交叉验证的方法求得。

Input: D, a dataset of ChunkSize from the incoming stream
　　$KMax$, the maximum number of classifiers
　　ENS, a set of previously trained classifiers
Output: ENS, a set of classifiers with updated weights
Train classifier f' from D;
Compute error rate of f' via cross validation on D;
For each classifier f_i
　　Compute error rate of f_i on D;
WeakestFirst Pruning ($KMax$);
$ENS \leftarrow ENS \cup \{f'\}$;
For each classifier f_i in ENS
　　Compute w_i;
Select subset from ENS in terms of Orientation based method.

图 1　OSEN 算法

随着数据的逐渐增加，集成中存在的基分类器个数也随之增加。这样既能增加存储空间的使用，又会降低预测速度，所以集成学习方法都会选择一种减枝策略来删除基分类器。本文采用 WeakestFirst 减枝策略，即设定一个 $KMax$ 值，当集成中基分类器个数大于 $KMax$ 时，使用新产生的基分类器取代泛化误差最大的基分类器。

算法中的权重计算方法如下：对于简单投票法(V) $w_i=1$；对于加权平均法(WV) $w_i=\log((1-e_i)/e_i)$，其中 e_i 表示集成中分类器 i 在数据集 D 上的泛化误差。

本文将选择性集成的思想引入对概念漂移数据流的挖掘算法中。根据基分类器的输出结果向量对基分类器进行评价，从当前所有基分类器中选择一个基分类器子集参与集成，以降低集成的泛化误差。

3 基于方向的基分类器选择方法

考虑一个包含 m 个实例的数据集 D。假定集成系统在数据集 D 上的期望输出为 $C=(c_1,c_2,\cdots,c_m)$，其中 c_i 表示第 i 个实例的期望输出。令 $F_j=(f_{j1},f_{j2},\cdots,f_{jm})(t=1,\cdots,T)$ 表示第 j 个基分类器的实际输出，其中 f_{ji} 表示第 j 个基分类器在第 i 个实例上的实际输出。

该基于结果向量方向的基分类器选择方法[11]的基本思想是：把基分类器对数据集 D 的分类输出看成一个 m 维向量，根据各基分类器在数据集 D 上的分类结果输出向量 F_j $(j=1,\cdots,n)$(n 为当前集成包含基分类器个数)计算出一个参考向量 S_{ref}。再分别计算向量 $F_j(j=1,\cdots,n)$ 与 S_{ref} 之间的夹角余弦值，选择夹角余弦值大于 0(夹角小于 $\pi/2$)的基分类器进行结果集成，即分配 0 权重给夹角大于 $\pi/2$ 的对应的基分类器。

设 $S_j=(s_{j1},s_{j2},\cdots,s_{jm})$ 表示集成中第 j 个基分类器对数据集 D 中实例分类正确与否的信号向量，定义为

$$s_{ji}=\begin{cases}1, & \text{if } f_j(x_i)=c_i\\ -1, & \text{if } f_j(x_i)\neq c_i\end{cases},i=1,2,\cdots,m$$

显然，当第 j 个基分类器在第 i 个实例上的实际输出正确时，$s_{ji}=1$，否则 $s_{ji}=-1$。

定义集成的平均信号向量为

$$\bar{S}=\frac{1}{n}\sum_{j=1}^{n}S_j,\text{即 }\bar{s}_i=\frac{1}{n}\sum_{j=1}^{n}s_{ji}$$

从上式不难看出，若 $\bar{s}_i$ 大于 0，表明集成中超过一半的基分类器对实例 x_i 分类正确。

定义参考向量 S_{ref} 的方向为第一象限主对角线在以 $\bar{S}$ 为法向量确定的平面上的投影。S_{ref} 按式(1)计算。

$$S_{ref}=\alpha o-\bar{S} \tag{1}$$

其中，o 表示沿第一象限对角线方向的向量 $o=(1,1,\cdots,1)$，α 为一常数，根据 S_{ref} 与 $\bar{S}$ 正交，可得 $\alpha=|\bar{S}|^2/o\cdot\bar{S}$。

例如，设一个由 4 个样本组成的数据集，当前集成在此数据集上的平均信号向量为 $\bar{S}=\{1/2,-1/2,0,1\}$，这说明，集成中有 75%的基分类器对第 1 个实例分类正确；25%的基分类器对第 2 个实例分类正确；50%的基分类器对第 3 个实例分类正确；对于第 4 个实例，所有基分类器分类结果都正确。根据式(1)可计算出 $\alpha=3/2$ 和 $S_{ref}=\{1,2,3/2,1/2\}$。

这种基分类器选择方法，在得到基分类器的输出结果后，通过式(1)可直接计算出参考向量，每个基分类器的输出结果向量只需与这个参考向量进行一次比较，因此这种基分类器选择方法的时间复杂度为 $O(n)$(n 为集成中基分类器个数)。

4 实验与分析

4.1 实验数据

根据概念漂移发生的频率和幅度可将其分为两类，突发性概念漂移和渐进性概念漂移。本文选取文献[12]中具有突发性概念漂移特性的数据集 SEA 和渐进性概念漂移特性的数据集 Hyperplane 进行算法性能的比较分析。

4.2 实验结果

本文实验设定 WeakestFirst 减枝策略的参数 $KMax$ 为 25。在每个时刻都需要一个验证集来确定各基分类器的泛化误差和权重，从而进行基分类器的减枝和选择。本文选择 t 时刻的训练集 D_t^v 作为 t 时刻的验证集 D_t^v。实验选择 naive Bayes 作为集成中的基分类器学习模型。

实验比较了 4 种集成策略：简单投票法集成(V)、加权平

均集成(WV)、应用 OSEN 的简单投票集成(OSEN/V)和加权平均集成(OSEN/WV)的分类性能。在进行结果集成时，方法 V 对集成中每个基分类器都给予相同的权重 1；方法 WV 根据各基分类器在验证集上的泛化误差来分配权重；方法 OSEN/V，即使用基于方向的基分类器选择方法，选择在验证集上的分类结果向量与参考向量之间夹角余弦小于 0 的基分类器赋予 0 权重，其余给予 1 权重；方法 OSEN/WV 使用同 OSEN/V，即使用基于方向的基分类器选择方法，选择一部分基分类器给予 0 权重，其余基分类器的权重设定方法同 WV。

4.2.1 在 SEA 上的实验结果

根据 SEA 当前时刻的概念使用事先抽取出的测试集 D_{te}^1，D_{te}^2，D_{te}^3 和 D_{te}^4 来验证当前集成的分类性能。图 2(a)和图 2(b)分别显示出了数据块尺寸为 500 和 1000 的分类准确性的比较结果。横坐标表示时刻，纵坐标表示分类预测准确性。从图中不难发现，OSEN/V 和 OSEN/WV 能快速收敛到新概念，分类性能明显优于 V 和 WV。在大多数情况下，OSEN/WV 的分类性能最好，发生概念漂移时，适应新概念速度最快。但是，在实例块大小为 500 的情况下，OSEN/WV 不太稳定，对应的曲线噪音较大。我们认为，发生这种现象的原因是，对验证集中的噪音数据过适应了。在实例块大小为 1000 的情况下，OSEN/WV 几乎在所有的时间段都表现出优于其他方法的性能。OSEN/V 的分类性能仅次于 OSEN/WV。V 和 WV 在实例块大小为 500 的情况下，性能相似，WV 比 V 略好。但在数据块大小为 1000 的情况下，V 的性能明显下降，主要原因是集成中保留了 25 个基分类器，并且对它们给出的结果给予相同的重视程度，导致过期概念对当前分类结果产生较大影响，所以分类效果特别差。

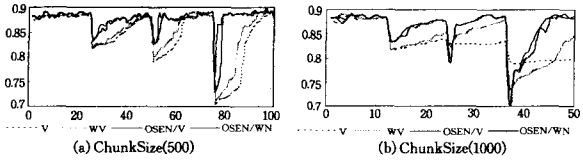


图 2 SEA

4.2.2 在 Rotating Hyperplane 上的实验结果

考虑 4 个不同的数据块尺寸：250，500，750 和 1000。测试数据集来自数据流本身，$t+1$ 时刻的数据块作为 t 时刻集成的测试数据集。

表 1 给出了 OSEN/WV 与 WV 在 9 个 Rotating Hyperplane 数据集上的比较结果。表中每一单元对应着一个 β 和 γ 取不同值的数据集。每个数字表示 OSEN/WV 与 WV 相比，其准确性百分比的提升。每个单元从上至下分别表示数据块尺寸取 250，500，750 和 1000 的比较结果。从表中不难发现，OSEN/WV 在所有情况下分类性能均优于 WV，并且概念漂移程度越大，性能提高越明显。

表 1 选择性集成在 Rotating Hyperplane 数据集上的准确性%提升

	γ=0.1	γ=0.5	γ=1
β=2	0.2	0.6	0.8
	1.1	1.0	1.1
	0.9	0.6	0.5
	1.1	0.6	0.5
β=5	0.2	2.2	3.5
	0.7	2.0	3.1
	0.1	1.9	3.5
	0.4	2.6	3.6
β=8	1.2	2.8	2.7
	1.7	2.9	3.2
	1.5	2.9	3.3
	1.3	3.1	3.3

4.3 error-ambiguity 分解

利用 error-ambiguity 分解对 OSEN 构建的 naive Bayes 集成在解决分类问题时的性能进行分析。error-ambiguity 分解是集成学习中一种重要的分析技术，它直接来源于 A. Krogh 和 J. Vedelsby 推导出的重要公式 $E=\bar{E}-\bar{A}$，这里 E 是集成的泛化误差，$\bar{E}$ 是集成中基分类器的平均泛化误差，$\bar{A}$ 表示集成中基分类器的平均 Ambiguity（在一定程度上可视为基分类器之间的差异度）。显而易见，$\bar{E}$ 越小，$\bar{A}$ 越大，集成效果越好。

本文拟从 Krogh 和 Vedelsby 的分解方法来解释 OSEN 的工作机制。分别对使用 OSEN 的简单投票法 OSEN/V 和未使用 OSEN 的简单投票法 V 进行了 error-ambiguity 分解实验比较。如 4.2 节所述，可得到从时刻 1 到时刻 T 集成的累积泛化误差 $E=\sum_{t=1}^{T}E_t$；计算 t 时刻参与集成的每个基分类器对当前测试数据集的误差，取平均值得到平均泛化误差 $\bar{E}_t$，累积平均泛化误差为 $\bar{E}=\sum_{t=1}^{T}\bar{E}_t$；再根据公式 $E=\bar{E}-\bar{A}$，可以计算出 $\bar{A}$。

图 3 给出了 OSEN/V 与 V 在数据集 SEA(500)和 Rotating Hyperplane(500，$K=8$，$T=1$)上的 error-ambiguity 分解。从图 3 不难发现，不使用选择性集成的简单投票结果集成法的累积平均泛化误差明显高于使用 OSEN 的简单投票结果集成法的累积平均泛化误差，所以 OSEN 算法成功的主要原因是它能够显著降低 $\bar{E}$。

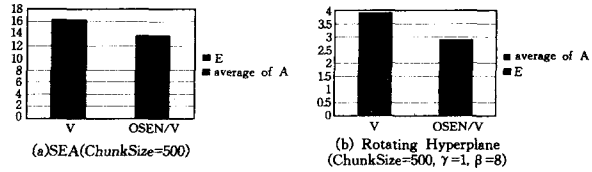


图 3 V 和 OSEN/V 两种集成方法的 error-ambiguity 分解

结束语 本文对基分类器结合集成方法进行了研究，提出了一种处理概念漂移数据流的选择性集成学习算法 OSEN。该算法把当前时间块数据作为验证集，根据各基分类器在验证集上的结果输出向量方向与参考向量方向之间的偏离程度，选择一个基分类器子集参与最终集成。实验结果表明，该算法能有效降低处理概念漂移数据流的泛化误差。本文使用 error-ambiguity 分解对 OSEN 构建的 naive Bayes 集成在解决分类问题时的性能进行了分析。实验研究说明，与简单投票法相比，OSEN 成功的主要原因是它显著降低了平均泛化误差。

事实上，集成学习技术除了用于监督学习和非监督学习，还用于多示例学习。这一方面说明集成学习的效用已经受到了广泛的认可，其适用范围正逐渐扩大；另一方面也说明集成学习的研究空间还很大，尤其是将集成学习技术应用到监督学习之外的场合，因此还有很多重要的问题需要研究。

参 考 文 献

[1] Klinkenberg R. Learning Drifting Concepts Example Selection vs Example Weighting[J]. In Intelligent Data Analysis (IDA), Special Issue on Incremental Learning Systems Capable of Dealing with Concept Drift, 2004, 8(3): 281-300

[2] Littlestone N, Warmuth M K. The weighted majority algorithm [J]. Information and Computation, 1994, 108: 212-261

[3] Freund Y, Schapire R. A decision-theoretic generalization of on-line learning and an application to boosting[J]. J. Computer and Sys. Sciences, 1997, 55: 119-139

[4] Street W N, Kim Y S. A streaming ensemble algorithm for large-scale classification [C] // Proceeding of the 7th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining. 2001: 377-382

[5] Kolter J Z, Maloof M A. Using additive expert ensembles to cope with concept drift[C]//Proceedings of the Twenty-second International Conference on Machine Learning. New York, NY: ACM Press, 2005: 449-456

[6] 孙岳,毛国君,等. 基于多分类器的数据流中的概念漂移挖掘[J]. 自动化学报, 2008, 34(1): 93-97

[7] Wang H, Fan W, Yu P S, et al. Mining concept - drifting data streams using ensemble classifiers[C]//Proc. 9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD-2003). ACM Press, 2003: 226-235

[8] Zhou Z H, Wu J X, Tang W. Ensembling neural networks: many could be better than all[J]. Artificial Intelligence, 2002, 17(1/2): 239-263

[9] Giacinto G, Roli F. Design of effective neural network ensembles for image classification purposes[J]. Image and Vision Computing, 2001, 19(9/10): 699-707

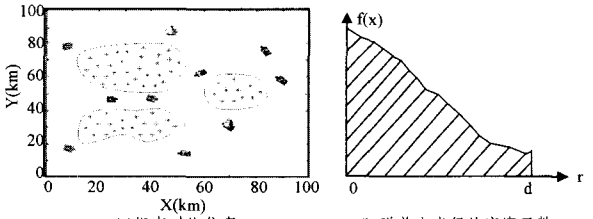
[10] Bakker B, Heskes T. Clustering ensembles of neural network models[J]. Neural Networks, 2003, 16: 261-269

[11] Martinez-Munoz G, Suarez A. Pruning in ordered bagging ensembles[C]//23rd International Conference in Machine Learning. ACM Press, 2006: 609-616

[12] Tsymbal A, Pechenizkiy M, Cunningham P, et al. Dynamic integration of classifiers for handling concept drift[J]. Information Fusion, 2008, 9(1): 56-68

(上接第 203 页)

δ 决定，r 为跑道落点距目标点的距离。不同的坦克可以选择一组不同的参数值，这样就可以让不同坦克的射击动作具有不同的水准，从而更加逼真地模拟现实中各坦克的射击情况。



(a)坦克对战仿真　　(b)弹着点半径的密度函数

图 5　虚拟坦克对战仿真

由实验可知：

1)坦克在攻击目标时，只能以一定的概率命中目标，而不是简单的命中或不命中；

2)通过调节对射击动作建模的正态区间参数，可以让不同的坦克具有不同的命中率，这样就可以模拟各种不同熟练程度的坦克乘员的射击行为。

结束语　1)基于不确定性空间的概率，明确了不确定性动作的具体定义；2)提出了概率区间的概念及其完备性问题，与代数区间相比，用概率区间对不确定性动作建模可以更加精确地刻画动作效应的分布情况；3)着重对正态区间进行了讨论，定义了正态区间、正态区间运算及其完备化结果。可以从如下两方面进一步研究：1)将概率区间扩展到二维，甚至更高维度；2)针对特定动作效应分布情况，定义具体的概率区间并研究相关性质。

参 考 文 献

[1] Wray R E, Laird J E, Nuxoll A, et al. Synthetic Adversaries for Urban Combat Training[C]//Proc. of the 2004 Innovative Applications of Artificial Intelligence Conference. San Jose, CA, July 2004

[2] Raza M, Sastry V V S S. Variability in Behavior of Command Agents with Human-Like Decision Making Strategies[C]// Proc. of Tenth International Conference on Computer Modeling and Simulation (uksim 2008). published by IEEE, 2008: 562-567

[3] Stytz M R. Progress and Prospects for the Development of Computer-generated Actors for Military Simulation: Part 1[J]. Introduction and Background, 2003, 12(3): 311-325

[4] Cox C, Fu D. AI for Automated Combatants in a Training Application[C]//Proc. of the Second Australasian Conference on Interactive Entertainment. Sydney, Australia: Creativity & Cognition Studios Press, 2005: 57-64

[5] Tian Z H, Zhao L, Jia Y. Research on Consistent Measurement of Uncertainty Based on Entropy[C]//Proc. of the International Conference on Intelligent Computation Technology and Automation. Changsha, China, Published by IEEE CS, 2008: 684-697

[6] Chery L H, Daniel P L. Interval methods for modeling uncertainty in RC timing analysis[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (0278-0070), 1992, 12(11): 1388-1401

[7] Liu B. Uncertainty Theory (2nd ed) [M]. Berlin, German: Springer-Verlag, 2007

[8] Funge J, Tu X, Terzopoulos D. Cognitive Modeling: Knowledge, Reasoning and Planning for Intelligent Characters[C]// Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. New York: ACM Press, 1999: 29-38

[9] Funge J. AI for Computer Games and Animation: A cognitive Modeling Approach (1-56881-103-9). USA Natick, MA: A. K. Peters Ltd., 1999