基于超网络和投影降维的高维数据流在线分类算法

茹 蓓
(新乡学院计算机与信息工程学院 河南 新乡 453003)

摘 要 为了提高高维数据流在线分类的准确率,设计一种基于超网络和投影降维的高维数据流在线分类算法。将高维数据流的特征子集建模为超网络模型,算法的学习目标是搜索最优的超边集合,选出判别能力强的特征子集。利用高斯核将高维空间的数据点投影到低维空间,采用梯度下降法计算数据点间的相似性矩阵。基于贝叶斯分类器模型更新机制,动态地学习新到达的数据流,基于学习的结果更新超网络的超边,再利用超网络指导分类器进行分类。仿真结果表明,该算法实现了较高的分类准确率,并且对于噪声也具有较好的鲁棒性。

关键词 超网络 超图 高维数据流 数据流分类 贝叶斯分类器 数据降维

中图分类号 TP391 **文献标志码** A DOI: 10.3969/j.issn.1000-386x.2020.10.044

ONLINE CLASSIFICATION ALGORITHM FOR HIGH DIMENSIONAL DATA STREAM BASED ON HYPERNETWORKS AND PROJECTION DIMENSION REDUCTION

Ru Bei
(*School of Computer and Information Engineering, Xinxiang University, Xinxiang 453003, Henan, China*)

Abstract To improve the classification accuracy of online classification of high dimensional data streams, this paper designs an online classification algorithm for high dimensional data stream based on hypernetworks and projection dimension reduction. It modeled the feature subsets of high dimensional data streams as hypernetworks, the learning objective is to search the optimal hyperedges sets and select the feature subsets with strong discriminant abilities. Gaussian kernel was used to project data points from high dimensional space to low-dimensional space, and gradient descent method was adopted to compute the similarity matrix of data points. Based on the model updating mechanism of Bayes classifier, the arrived data stream was learned dynamically. The hyperedges of hypernetworks was updated based on the learning results, and then the hypernetworks were used to guide the classifier for classification. The simulation experimental results show that this algorithm achieves high classification accuracy and has good robustness to noise.

Keywords Hypernetworks Hypergraphs High dimensional data stream Data stream classification Bayes classifier Data dimension reduction

0 引 言

近年来出现了许多数据流的实时分类应用,例如:金融市场通过对证券股指、期货数据进行实时分类和预测,能够快速察觉市场行情的变化,有助于提高投资收益和降低风险[1];社交媒体用户每天需要发布大量的图像数据流,通过对图像进行实时分类和预测,能够快速检测出非法图像[2]。对数据流的实时分类和预测存在巨大的应用价值[3],而图像、文档等高维数据是数据流的一个重要组成部分,其高维特性严重影响了分类器的计算效率和分类性能,成为实时数据流分类的一个难点[4]。

增量学习思想是当前数据流实时分类的一个重要手段,文献[5]提出了噪声消除的增量学习分类器,使用互信息近邻来检测噪声样本,通过增量学习检测数据流的类标签。文献[6]提出基于样本不确定性选择策略的增量数据流分类模型,从相邻训练集中按照样

收稿日期:2019-07-07。河南省软科学研究计划项目(192400410045)。茹蓓,教授,主研领域:信息处理,大数据。

本不确定性值选出“富信息”样本代表新概念样本集。增量学习主要通过一些评价指标检测出某些判别能力强的新到达样本,然后结合这些样本对分类器进行更新,此类方法主要以上一个时间窗口的模型为基础,导致变化剧烈窗口的训练集存在明显的偏差,难以实现理想的分类准确率。文献[7]将回归系统应用于时变环境,实现了回归神经网络,该方法通过回归模型主动学习神经网络的模型参数,实现了较为理想的性能,虽然采用了简单的神经网络结构,但该模型的回归模型依然存在训练时间长的问题。

基于神经网络的数据流分类器一般通过增量学习机制动态地更新神经网络的参数,采用多层神经网络处理高维数据流才能获得更好的效果,但多层神经网络的参数量较多,实时学习的难度较大[8]。本文从一个新的角度出发,对高维数据流进行分析和分类处理。采用超网络建模高维数据流的数据空间,利用基于高斯核的投影技术将高维数据投影到低维空间。利用贝叶斯模型的先验分布、后验分布和似然信息拟合数据流的动态特性,设计了贝叶斯模型的实时更新方法,实现了对高维数据流的实时分类处理。

1　高阶模型的超网络

超图[9]是离散数学中有限集合的子系统结构,超图的边为高阶边,称为超边,超边连接两个以上的顶点。将超图表示为 $G=(V,E)$,其中 V 和 E 分别是顶点集和超边集。一条超边是 V 的一个子集,超边的权重满足 $w(e)\geqslant 0$。顶点的度和超边的度 $d(v)$ 和 $\delta(e)$ 分别定义为:

$$d(v)=\sum_{e\in E}w(e)h(v,e) \tag{1}$$

$$\delta(e)=|e| \tag{2}$$

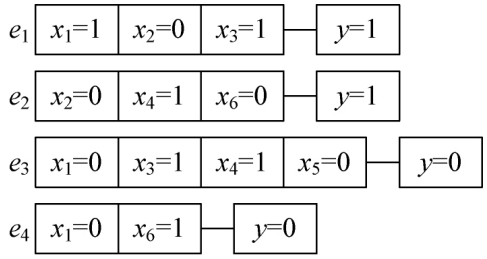
式中: $|e|$ 为边 e 的基数,将度等于 k 的超边简称为 k-超边。超边的度值越高,则该超边的模式判别能力越强。如果一个超图的超边均为 k-超边,则称其为 k-均匀超图,所以 2-均匀超图即为传统意义的图,3-均匀超图即为三元组的集合,以此类推。

超网络[10]是超图结构的高阶表示,超网络的顶点定义为一个变量及其取值,超边定义为顶点间的高阶连接,超边的权重表示连接的强度。超网络是大量超边的集合,可表征高维数据特征之间的高阶关系。

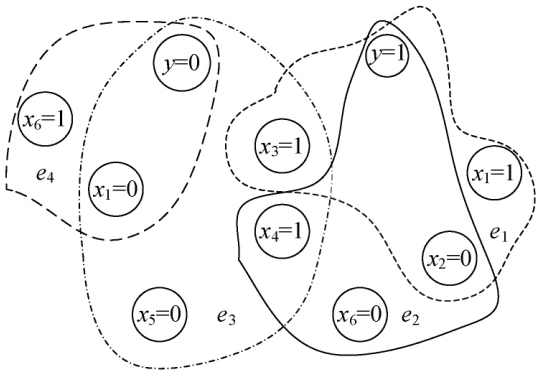
将超网络定义为一个三元组形式 $H=(V,E,W)$,其中 W 表示超边的权重集。超边是两个以上顶点的集合,表示为:

$$e_i=\{v_{i1},v_{i2},v_{i3},\cdots,v_{i|e_i|},y_i\} \tag{3}$$

式中: y_i 表示第 i 个超边 e_i 的类标签。超边的权重反映了超边对于类标签的判别能力,所以可将超边考虑为一个弱学习机,学习分类器所需的特征子集。图 1 为一个超网络及超图结构的示意图。



(a) 超网络实例



(b) 超图实例

图 1　超网络和超图的示意图

给定一个数据点集 $x^{(n)}$、一个类标签集 Y 和一个超网络 H,选择 $x^{(n)}$ 超边加权调和值最大的标签,将该极大值标签作为 $x^{(n)}$ 的类标签。超网络模型的分类步骤如下:

步骤 1　计算超边集 E 中所有 $y\in Y$ 超边的权重之和:

$$\tilde{w}_y=\sum_{i=1}^{|E|}\{w(e_i)f(x^{(n)},e_i)\varphi(y,y_i)\} \tag{4}$$

式中: $w(e_i)$ 表示 e_i 的权重。

步骤 2　选择总权重最大的标签作为 $x^{(n)}$ 的标签:

$$\hat{y}(n)=\arg\max_{y\in Y}\tilde{w}_y \tag{5}$$

定义两个指示函数 $f(x^{(n)},e_i)$ 和 $\varphi(y^{(n)},y_i)$,如果 e_i 匹配 $x^{(n)}$,则 $f(x^{(n)},e_i)=1$; 如果 $y^{(n)}=y_i$,则 $\varphi(y^{(n)},y_i)=1$,即:

$$f_i^{(n)}=f(x^{(n)},e_i)=\begin{cases}1 & \exp\{c(x^{(n)},e_i)-\delta(e_i)\}>\theta\\ 0 & \text{其他}\end{cases} \tag{6}$$

$$\varphi_i^{(n)}=\varphi(y^{(n)},y_i)=\begin{cases}1 & y^{(n)}=y_i\\ 0 & \text{其他}\end{cases} \tag{7}$$

式中: $c(x^{(n)},e_i)$ 表示匹配的数量; θ 为匹配的阈值,用

于增强对噪声数据的鲁棒性。

超网络的模型结构对分类算法的性能具有极大的影响，学习程序的目标是寻找最优的超边集，即从一个特征集选出特征子集。该问题主要分为 3 个子问题：(1) 选择构建超边的变量子集；(2) 确定超边的度；(3) 确定模型的超边数量。子问题(1)和(2)影响分类器的分类性能，子问题(2)和(3)影响模型的计算复杂度。如果特征为二进制值，那么特征集的规模为 $2^{2^{|x|}}$，$|x|$ 为数据的维度，超网络模型对于高维数据的计算复杂度较高，因此为超网络设计一个高效的模型学习方法，即基于贝叶斯模型的超网络更新机制。

2 基于贝叶斯模型更新超网络

图 2 为超网络数据流分类算法的结构框图，使用贝叶斯更新方法学习超网络的结构，学习的内容包括生成超边、更新超边权重和评估超网络模型。首先，从训练数据集提取超边，构建初始化超网络，将超边和训练数据匹配计算超边的权重。然后，将训练集分类估算模型的适应度，将低权重超边替换为新生成的超边，动态地修改模型。

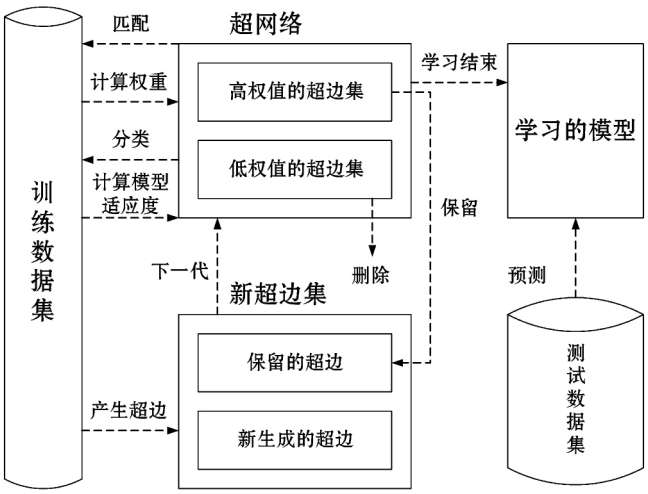


图 2 超网络高维分类算法的结构框图

2.1 贝叶斯模型的学习方法

引入贝叶斯模型学习超网络，贝叶斯模型假设后验分布和先验分布分别为当前种群(当前超网络)和上一代种群(上一代超网络)。模型的适应度定义为后验概率，适应度同时反映了数据的判别能力和模型的复杂度。

设 H_t 是第 t 次迭代的超网络。设数据集为 $D=\{X,Y\}$，数据量为 $X=\{x^{(n)}\}_{n=1}^{N}$，类标签集为 $Y=\{y^{(n)}\}_{n=1}^{N}$，分类问题定义为贝叶斯法则的条件概率：

$$p(H_t\mid X,Y)=\frac{p(Y\mid X,H_t)\,p(H_t\mid X)}{p(Y\mid X)} \tag{8}$$

式中：$p(Y\mid X,H_t)$ 和 $p(H_t\mid X)$ 分别为似然信息和先验分布；$p(Y\mid X)$ 是一个正则常量。后验分布关于似然信息和先验分布的乘积成正比例关系：

$$p(H_t\mid X,Y)\propto p(Y\mid X,H_t)\,p(H_t\mid X) \tag{9}$$

定义 H_t 的适应度函数 F_t 为后验分布的对数，目标函数变为最大化式(10)：

$$F_t=\log p(Y\mid X,H_t)+\log p(H_t\mid X) \tag{10}$$

即

$$H^{*}=\arg\max_{H_t}F_t \tag{11}$$

2.2 采用贝叶斯模型学习超网络结构

计算超网络的适应度需要定义模型的先验信息和似然信息，将经验先验分布 $p(H_t\mid X)$ 定义为目标问题的先验知识。超网络的先验信息包括两点：(1) 变量和类标签之间的相似性矩阵，采用基于投影的相似性矩阵选择数据，产生初始化超边；(2) 缩小模型的规模，从上一次的后验分布 $p(H_{t-1}\mid Y,X)$ 计算当前迭代的经验先验分布：

$$p(H_t\mid X)\propto p(H_{t-1}\mid Y,X) \tag{12}$$

$$p(H_t\mid X)\propto\frac{1}{|H_{t-1}|}\prod_{e\in E_n}P(e)\approx\frac{1}{|H_{t-1}|}\prod_{e\in E_n}\prod_{x_i\in e}P_1(x_i) \tag{13}$$

$$\text{s.t.}\quad |H_t|=\sum_{e\in E_t}\delta(e),\ E_n=E_t-E_{t-1}$$

式中：E_t 为 H_t 的超边集；$P(e)$ 表示超边 e 的产生概率；$P_1(x_i)$ 表示选择变量 x_i 的概率；$P_1(x_i)$ 关于 x_i 和类标签之间的相似性成正比例关系。仅需要对每个新到达时间窗口计算一次 $P_1(x_i)$，在更新过程中无需改变该概率值。通过上述方法可确定超网络模型的 3 个主要参数：超边包含的变量、超边的度和超边数量。

似然定义为从 H_t 正确分类 Y 的条件概率，X 表示了 H_t 的判别能力。统计模型对训练数据正确匹配和错误匹配的数量，再计算每个超边的加权调和值，通过评估超边的判别能力估算似然信息。如果超边和一个数据实例的标签匹配，则认为该超边正确匹配，否则匹配失败。总似然计算为各个数据实例似然的乘积：

$$p(Y\mid X,H_t)=\prod_{n=1}^{N}p(y^{(n)}\mid x^{(n)},H_t) \tag{14}$$

将经验似然定义为：

$$p(y^{(n)}\mid x^{(n)},H_t)\equiv\frac{\sum_{i=1}^{|E_t|}w(e_i)\{f_i^{(n)}\cdot\varphi_i^{(n)}-f_i^{(n)}\cdot(1-\varphi_i^{(n)})\}}{\sum_{i=1}^{|E_t|}w(e_i)}=$$

$$\frac{\sum_{i=1}^{|E_t|} w(e_i)\{f_i^{(n)} \cdot (2\varphi_i^{(n)} - 1)\}}{\sum_{i=1}^{|E_t|} w(e_i)} \tag{15}$$

式中: $w(e_i)$ 为第 i 个超边的权重; E_t 为 H_t 的超边集合。如果经验似然小于 0,则将其设为一个小的正数来防止出现负似然的情况。可获得:

$$p(Y \mid X, H_t) = \prod_{n=1}^{N}\left[\frac{\sum_{i=1}^{|E_t|} w(e_i)\{f_i^{(n)} \cdot (2\varphi_i^{(n)} - 1)\}}{\sum_{i=1}^{|E_t|} w(e_i)}\right] \tag{16}$$

s. t. $f_i^{(n)} = f(x^{(n)}, e_i)$, $\varphi_i^{(n)} = \varphi(y^{(n)}, y_i)$

在训练程序中,如果正确匹配一条超边,那么 $f_i^{(n)} \cdot \varphi_i^{(n)}$ 和 $f_i^{(n)}(1 - \varphi_i^{(n)})$ 分别等于 1 和 0; 如果错误匹配,则分别为 0 和 1; 如果匹配失败,则两个值都为 0。

超边的权重是关于正确匹配情况和错误匹配情况的函数:

$$w(e_i) = \alpha\sum_{n=1}^{N} f_i^{(n)} \cdot \varphi_i^{(n)} - (1 - \alpha)\sum_{n=1}^{N} f_i^{(n)} \cdot (1 - \varphi_i^{(n)}) + \frac{\beta}{|e_i|} = \sum_{n=1}^{N} f_i^{(n)} \cdot \varphi_i^{(n)} - (1 - \alpha)\sum_{n=1}^{N} f_i^{(n)} + \frac{\beta}{|e_i|} \tag{17}$$

式中: α 是小于 1 的常量,用来增加正确预测的概率; β 是一个小的正常量。因为度小的超边匹配率更高,所以如果两个超边的度接近,则选择度小的超边,β 不仅减少了模型的复杂度,而且增加了模型的泛化效果。如果 $w(e)$ 小于 0,则设为 0,防止出现负权重。

结合式(13)的经验先验和式(16)估算的似然,将超网络的适应度修改为:

$$F_t = \log p(Y \mid X, H_t) + \log p(H_t \mid X) \approx \sum_{n=1}^{N} \log \frac{\sum_{i=1}^{|E_t|}\{w(e_i) f_i^{(n)}(2\varphi_i^{(n)} - 1)\}}{\sum_{i=1}^{|E_t|} w(e_i)} + \lambda \frac{|H_0|}{|H_t|} + \zeta\sum_{e \in E_t} \log \sum_{x_i \in e} P_1(x_i) \tag{18}$$

式中: λ 和 ζ 均为正常量,λ 负责调节模型的大小,ζ 负责调节先验信息的变量选择能力。

综上所述,提高适应度的措施有三点: (1) 每次迭代保留权值高的超边; (2) 选择 $P_1(x)$ 概率大的变量生成超边; (3) 在保持模型判别能力的情况下,尽量减少超边的数量。超网络更新的结束条件是在连续几次迭代后,F_t 均不大于 $F_{\max}$,或者达到最大迭代次数 $I_{\max}$。

2.3　基于高斯核投影的相似性度量

直接度量高维数据的相似性不仅准确率低而且计算复杂度也较高,因此本文采用高斯核的投影技术计算高维数据之间的相似性矩阵。设 $S(x_i, x_j)$ 为数据 x_i 和 x_j 间的相似性,其投影矩阵定义为 $[P]_{i,j} = S(f_{\mathrm{DR}}(x_i), f_{\mathrm{DR}}(x_j))$,其中: i 和 j 是矩阵的行和列,f_{DR} 为投影函数。本文的目标是将高维空间的数据点投影到低维空间,设目标相似性矩阵是一个方阵 T,则需要优化以下的目标函数:

$$J_s = \frac{1}{2\|M\|_1}\sum_{i \neq j}^{N} M_{i,j}(P_{i,j} - T_{i,j})^2 \tag{19}$$

式中: M 为目标相似性的权重矩阵; $\|M\|_1 = \sum_{i=1}^{\zeta}\sum_{j=1}^{\omega}|M_{i,j}|$ 表示矩阵 M 的 l_1 范数。如果每对数据点超过了目标相似性,此时的目标函数则是最小值。

使用高斯核定义投影点之间的相似性[11],即 $S(x_i, x_j) = \exp(-\|x_i - x_j\|_2^2/\sigma_{\mathrm{p}})$,其中 σ_{p} 为缩放因子。相似性矩阵 P 中的元素可定义为:

$$P_{i,j} = \exp(-\|f_{\mathrm{DR}}(x_i) - f_{\mathrm{DR}}(x_j)\|_2^2/\sigma_p) \tag{20}$$

设 $c(x)$ 为低维空间的相似性度量(例如: 欧氏距离),可将目标矩阵 T 中的元素定义为:

$$T_{i,j} = \exp\left(-\frac{\|c(x_i) - c(x_j)\|_2^2}{\sigma_c}\right) \tag{21}$$

式中: σ_{c} 为缩放因子。

采用核方法实现投影处理,设 $\boldsymbol{\Phi} = \phi(X)$ 为 Hilbert 高维空间[12]的数据矩阵,$\phi(X)$ 为高维空间的投影函数,然后学习从高维空间到低维空间的线性映射。将矩阵 W 重新定义为:

$$W = \phi(X)^{\mathrm{T}} A = \Phi^{\mathrm{T}} A \tag{22}$$

式中: $A \in \mathbf{R}^{n \times m}$ 为系数矩阵。将投影定义为数据点的线性组合,可获得以下的投影方法:

$$Y^{\mathrm{T}} = W^{\mathrm{T}}\Phi^{\mathrm{T}} = A^{\mathrm{T}}\Phi\Phi^{\mathrm{T}} = A^{\mathrm{T}}K \tag{23}$$

式中: $K = \Phi\Phi^{\mathrm{T}} \in \mathbf{R}^{n \times n}$ 是数据的核矩阵,表示 Hilbert 空间数据点的内积,即 $K_{ij} = \phi(x_i)^{\mathrm{T}}\phi(x_j)^{\mathrm{T}}$。

核方法学习系数矩阵 A,采用梯度下降法优化目标:

$$\frac{\partial J}{\partial A_{k,t}} = \frac{1}{\|M\|_1}\sum_{i=1}^{N}\sum_{j=1}^{N} M_{i,j}(P_{i,j} - T_{i,j})\frac{\partial P_{i,j}}{\partial A_{k,t}} \tag{24}$$

$$\frac{\partial P_{i,j}}{\partial A_{k,t}} = -\frac{2}{\sigma_{\mathrm{P}}}P_{i,j}(Y_{i,t} - Y_{j,t})(K_{i,k} - K_{j,k}) \tag{25}$$

2.4　学习超网络的方法设计

贝叶斯更新模型包括: 生成超边、选择超边和更新超边。生成超边需要确定两个属性: (1) 超边的数据变量; (2) 超边的度。图 3 为生成超边的示意图。

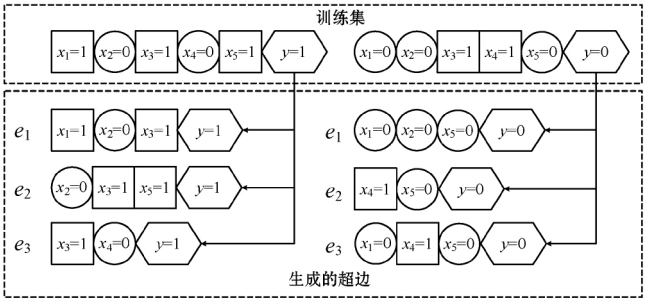


图 3 生成超边的示意图

根据变量之间的投影相似性矩阵选择相似的变量，相似的变量组成一条超边。式(13)的 $P_1(x_i)$ 计算为:

$$P_1(x_i)=\frac{I(x_i;y)+\eta}{\sum_{j=1}^{|x|}\{I(x_j;y)+\eta\}} \tag{26}$$

式中: $I(x_i;y)$ 表示变量 x_i 和类 y 间的相似性(基于投影计算高维数据的相似性); η 为一个非负常量，该常量防止相似性过小。将 η 和式(18)的 ζ 设为反比例关系。

根据 H_{t-1} 超边度的概率决定 H_t 超边的度:

$$P(\delta(e)=K)=\frac{|E_{t-1}^{K}|+\varepsilon}{\sum_{k=K_{\min}}^{K_{\max}}(|E_{t-1}^{k}|+\varepsilon)} \tag{27}$$

式中: E_{t-1}^{k} 表示在迭代 $t-1$ 的 k-超边集; ε 为一个平滑常量因子。如果两个超边的判别能力接近，那么模型偏向于选择度数低的超边。生成超边的程序如算法 1 所示。

算法 1 生成超边的程序

输入: 数据集 $S=\{S_1,S_2,\cdots,S_N\}$。

输出: 超边 e。

/* 初始化程序* /

n←从 $|D|$ 随机选择数据样本;

d←$D[n]$;

K←通过式(27)计算边的度;

For each k from 1 to K do

　　idx←式(26)选择变量;

　　val←d 数据的 x_{idx};

　　v←(idx,val);

　　e←$e\cup\{y^{(n)}\}$;

end for

e←$e\cup\{y^{(n)}\}$;

$w(e)$←0;

在迭代过程中删除低权重的超边，增加新的超边，但可能存在相似的冗余超边，导致计算复杂度升高。因此式(18)的适应度有效维护了超网络的大小，如果 F_t 大于 F_{t-1}，则缩小超网络，否则扩大超网络。

设 R_t 和 G_t 分别表示在第 t 次迭代删除的超边数量和新产生的超边数量。将 R_t 定义为一个关于 t 的函数，G_t 定义为一个关于 R_t 的函数，R_t 和 G_t 分别定义为:

$$R_t=\frac{R_{\max}-R_{\min}}{\exp(t/\kappa)}+R_{\min} \tag{28}$$

$$G_t=\gamma_t\cdot R_t \tag{29}$$

式中: $R_{\max}$ 和 $R_{\min}$ 分别表示 R_t 值的上界和下界; κ 为一个常量，控制 $R_{\max}$ 到 $R_{\min}$ 的变化速率; γ_t 为泛化模型的比例系数，定义为:

$$\gamma_t=\begin{cases}\left(\frac{F_{t-1}}{F_t}\right)^{\tau} & \Delta F_t\geqslant 0\\ \frac{F_{\max}}{F_t} & \Delta F_t<0\end{cases} \tag{30}$$

式中: $\Delta F_t=F_t-F_{t-1}$; τ 控制种群缩小的速度，如果 $\tau=0$，那么种群规模保持不变。

图 4 为基于贝叶斯的超边更新算法流程图，根据数据集的先验知识生成初始化超网络，将超标权重和特征的判别能力作为贝叶斯的似然信息，将超网络的复杂度和适应度作为贝叶斯的后验分布，贝叶斯迭代地学习最优的超网络结构。基于贝叶斯的超边更新算法伪代码如算法 2 所示。

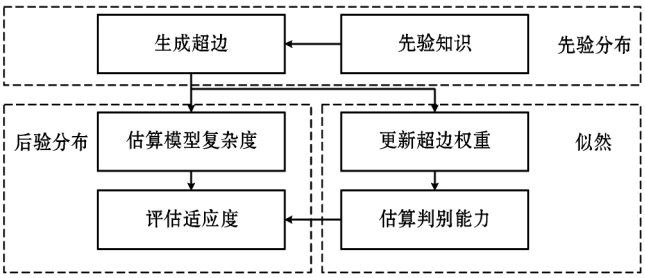


图 4 基于贝叶斯的超边更新算法流程示意图

算法 2 基于贝叶斯的超边更新程序

输入: 训练集 D，所有变量的相似性矩阵 SM，第 t 次迭代删除的超边数量 R_t，第 t 次迭代新生成的超边数量 G_t，第 t 次迭代的超边集 E_t，第 t 次迭代的超网络 H_t，最大迭代次数 $I_{\max}$，H_0 的初始化超边数量 H_{init}。

1. E_0←NULL，H_0←NULL; //初始化
2. 计算 $SM(D,SM)$; //式(26)
3. for each i from 1 to H_{init} do //建立初始化模型
4. 　　e_i←生成超边; //式(27)
5. 　　w_i←评价权重; //式(17)
6. 　　E_0←$E_0\cup e_i$;
7. end for
8. 利用式(18)评估当前模型的适应度;
9. E_1←利用式(22)删除低权重超边;
10. for each t from 1 to T do //超边迭代更新模块
11. 　　for i from 1 to G_t do

12. e_i←生成超边; //式(27)
13. w_i←评价权重; //式(17)
14. $E_0 \leftarrow E_0 \cup e_i$;
15. end for
16. 以式(18)评估当前模型的适应度;
17. if 未满足结束条件 then
18. $H \leftarrow H \times H_i$;
19. eles
20. break;
21. end if
22. E_{t+1}←删除低权重超边; //式(28)
23. 数据流分类

3 实　验

3.1 实验环境和实验方法

实验环境为 PC 机,处理器为 Intel Core i5-4570,内存为 16 GB,操作系统为 Ubuntu 14.04 LTS,软件的编译环境为 MATLAB R2017a。

将数据集排列成序列形式,每个到达数据首先用来测试在线分类器的分类性能,再用来更新分类器的模型。每组实验将每个数据集做 10 次置乱处理,置乱数据集作为输入数据,独立完成 10 次实验,统计 10 次实验的平均误差率和标准偏差值。

3.2 实验数据集

采用 10 个 UCI 高维数据集作为实验的 benchmark 数据集,表 1 为 benchmark 数据集的基本属性。数据集排列成序列形式来模拟数据流。

表 1　10 个高维 benchmark 数据集的基本属性

序号	数据集	属性	分类	样本量
1	Biodeg	41	2	1 055
2	Libras	90	15	360
3	Marketing	13	9	6 876
4	Optdigits	64	10	5 620
5	Penbased	16	10	10 992
6	Ring	20	2	7 400
7	Satimage	36	6	6 435
8	Sonar	60	2	208
9	Spambase	57	2	4 601
10	Texture	40	10	5 500

3.3 对比实验方法

选择 4 个近期的数据流在线分类器作为对比方法:

(1) 基于演化的在线贝叶斯分类器[13],简称为 OnlineBayes,该方法与本算法同属于采用贝叶斯分类器的在线分类方案。

(2) 基于元神经元的脉冲神经网络分类器[14],简称为 OnlineSpikingNeural,该方法是基于神经网络的在线分类器。

(3) 基于混合分类回归树和模糊自适应谐振网络的在线分类器[15],简称为 FAM-CART,该方法较为新颖,且性能较为突出。

(4) 在线的广义分类器[16],简称为 onlineuniversal,该分类器支持多种类型的数据流。

本文算法简称为 HypernetworksBayes。

3.4 分类器性能实验结果

1) 高维数据流的分类错误率。首先评估 5 个分类器对于 10 个高维数据流的分类准确率,采用分类误差率作为度量指标,图 5 为总体的统计结果。可以看出,HypernetworksBayes 对于 Biodeg 和 Spambase 2 个数据集的分类错误率高于 FAM-CART,HypernetworksBayes 对于 Ring 数据集的分类错误率也高于 onlineuniversal 算法,但结果也较为接近。对于数据流的二分类问题,本文算法与其他优秀的分类器较为接近,但对于其他高维数据的多分类问题,本文算法的优势则较为明显。FAM-CART 对于 10 个高维数据流均表现出较低的分类错误率,FAM-CART 采用分类回归树和模糊自适应谐振网络 2 项技术,模糊自适应谐振网络具有较强的自适应和自调节能力,而分类回归树对于连续性数据流具有较强的处理能力,因此实现了较好的效果。比较 OnlineBayes 和 HypernetworksBayes 可见,HypernetworksBayes 的分类错误率远低于 OnlineBayes 分类器,即本文的超图理论和基于高斯核投影的降维处理有效地提高了分类器的性能。总体而言,本文算法的分类错误率低于其他 4 种算法。

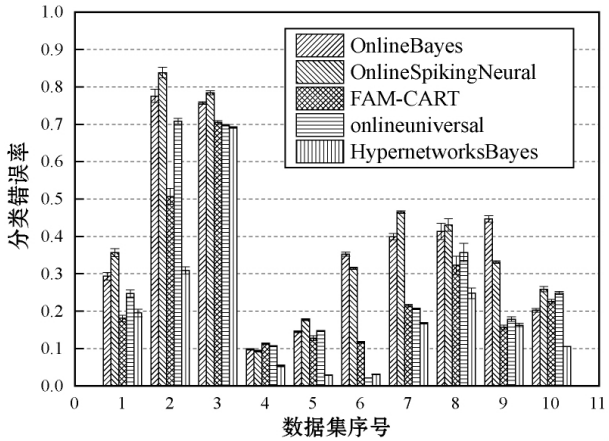


图 5　5 个分类器对高维数据流的分类错误率的比较

2) 分类器统计检验。通过 Friedman 检验[17]测试分类器对于 10 个数据集分类结果的差异，Friedman 检验的显著性设为 0.05。Friedman 检验获得的排名如图 6 所示。本文算法排名第一，且远高于第二名，再次验证了 HypernetworksBayes 分类器的性能优势。

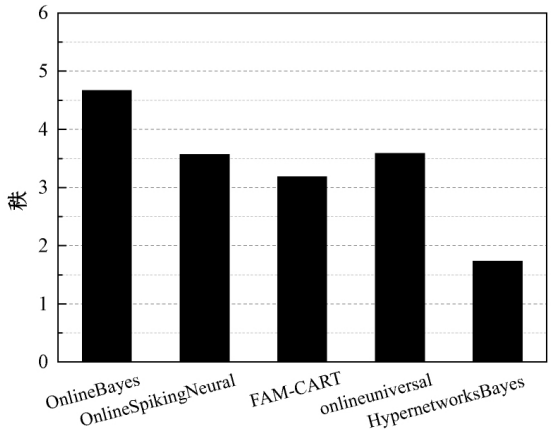


图 6 分类器的 Friedman 检验结果

Friedman 检验能够比较多个分类器之间的性能，而验后比较检验(post-hoc test) 能够详细比较多个分类器和单一分类器的性能。通过验后比较检验进一步分析分类器的性能，根据文献[18 - 19]的讨论，Hommel 检验适合评估数据流的在线分类器的性能，所以采用 hommel post-hoc 对分类器进行验后比较检验分析，结果如表 2 所示。观察表中各个分类器的 p-value 值，HypernetworksBayes 明显优于其他 4 个分类器，而对比方案中 OnlineSpikingNeural 和 FAM-CART 的性能也较为理想。

表 2 分类器的 hommel post-hoc 检验

在线分类器	p-value
OnlineBayes	8.83E - 12
OnlineSpikingNeural	9.54E - 03
FAM-CART	9.64E - 03
onlineuniversal	9.33E - 03

3.5 分类器的时间效率

时间效率是数据流在线分类器的一个重要性能指标，直接决定了分类器的应用价值。为了评估在线分类器的时间效率，统计了 5 个分类器每组实验分类程序的平均时间，结果如图 7 所示。可以看出，OnlineSpikingNeural 和 FAM-CART 的处理时间远高于其他 3 个算法，这 2 个算法均为基于神经网络的分类器，在训练神经网络的过程中难以同时在网络性能和时间效率两方面均取得理想的平衡。受益于贝叶斯模型的高计算效率，OnlineBayes 的时间效率最高。Onlineuniversal 分类器不包含复杂的计算和模型，也实现了极高的时间效率。HypernetworksBayes 的时间效率略低于 OnlineBayes 分类器和 Onlineuniversal 分类器，但是也足以满足处理实时数据流的时间需求。

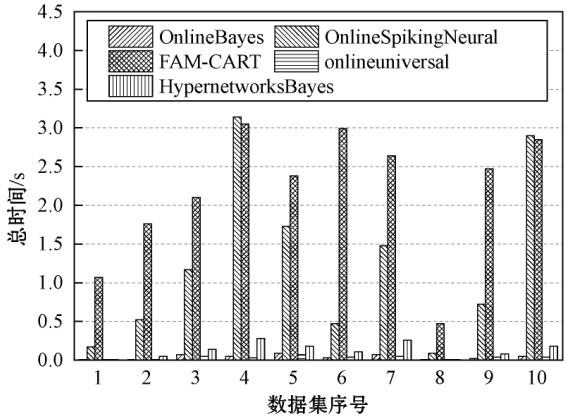
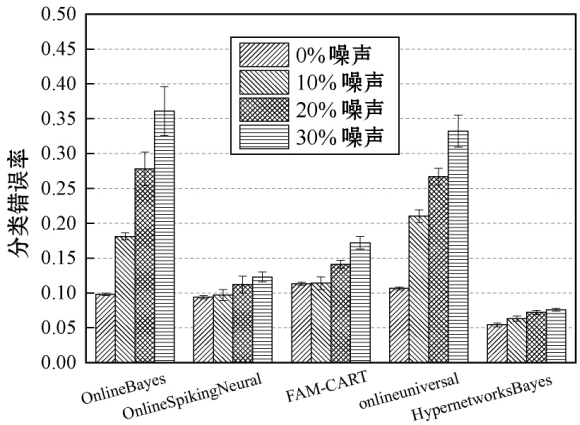


图 7 5 个在线分类器的平均处理时间的比较

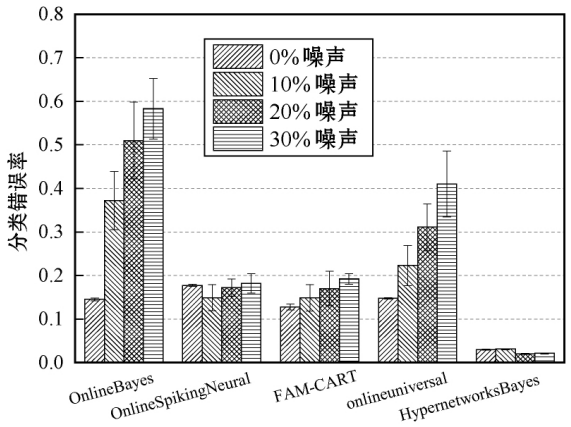
3.6 噪声数据流的分类实验结果

在实际的数据流应用中，数据流均伴随着不可忽略的噪声数据，因此测试了在线分类器对于噪声数据流的性能。随机选择一个错误类标签，替换训练数据集的正确标签，训练集的噪声数据比例分别设为 10%、20% 和 30%。训练集为噪声数据集，测试数据集为无噪声数据集，使用 10 折交叉验证完成每组数据集的实验，计算 10 次独立实验分类错误率的平均值。因为 Optdigits 和 Penbased 两个高维数据集的分类错误率较低，所以重点测试并统计了噪声数据对于 Optdigits 和 Penbased 两个高维数据集的影响。

图 8 为 Optdigits 和 Penbased 两个高维数据集的实验结果。可以看出，随着噪声比例的增加，本文算法分类错误率略有升高，但上升幅度较小。OnlineBayes 和 onlineuniversal 两个分类器受噪声的影响较大，几乎不具备噪声鲁棒性，因此可得结论: 本文的改进措施有效地增强了分类器的鲁棒性。



(a) Optdigits 数据集



(b) Penbased 数据集

图 8　噪声数据流的分类实验结果

4　结　语

本文设计了基于超网络和投影降维的高维数据流在线分类算法,利用贝叶斯模型的先验分布、后验分布和似然信息拟合数据流的动态特性,设计了贝叶斯模型的实时更新方法,实现了对高维数据流的实时分类处理。本文算法利用高斯核将高维空间的数据点投影到低维空间,实现了较高的高维数据流分类准确率,并且对于噪声也具有较好的鲁棒性。

本文模型依然存在一些限制: 无法处理不平衡数据流,而不平衡数据流也是实时数据流领域的一个细分领域; 仅支持数据维度权重相等的情况。未来将尝试引入分级的超网络模型,解决特征维度权重不相等的问题,从而可将本方案运用到推荐系统等多属性度量的应用中。

参 考 文 献

[1] 刘翼,嵩天,廖乐健. 基于时序流的移动流量实时分类方法[J]. 北京理工大学学报,2018,38(5):537 - 544.

[2] 杨融泽,柳毅. 面向异常数据流的多分类器选择集成方法[J]. 计算机工程与应用,2018,54(2):107 - 113.

[3] 张军,胡震波,朱新山. 基于 AdaBoost 分类器的实时交通事故预测[J]. 计算机应用,2017,37(1):284 - 288.

[4] 汪星,黄小瑜,刘瑄璞,等. 面向工业大数据的多层增量特征提取方法[J]. 西安电子科技大学学报(自然科学版),2018,54(4):106 - 111.

[5] Liu S M, Wang Z Q, Liu T, et al. Research on dynamic data streams classification with noise elimination using mutual nearest neighbor [J]. Journal of Frontiers of Computer Science and Technology,2016,10(1):36 - 42.

[6] 刘三民,孙知信,刘涛. 基于样本不确定性的增量式数据流分类研究[J]. 小型微型计算机系统,2015,36(2):193 - 196.

[7] Duda P, Jaworski M, Rutkowski L. Convergent time-varying regression models for data streams: tracking concept drift by the recursive parzen-based generalized regression neural networks [J]. International Journal of Neural Systems, 2018,28(2):1750048.

[8] Ali M, Anjum A, Yaseen M U, et al. Edge enhanced deep learning system for large-scale video stream analytics [C]//2008 IEEE International Conference on Fog and Edge Computing. IEEE,2018.

[9] Zhang L M, Gao Y, Hong C Q, et al. Feature correlation hypergraph: exploiting high-order potentials for multimodal recognition [J]. IEEE Transactions on Cybernetics,2017,44(8): 1408 - 1419.

[10] Baytas I M, Xiao C, Wang F, et al. Heterogeneous hypernetwork embedding [C]//2018 IEEE International Conference on Data Mining (ICDM). IEEE,2018:875 - 880.

[11] 钱美旋,叶东毅. 利用一维投影分析的无参数多密度聚类算法[J]. 小型微型计算机系统,2013,34(8):1866 - 1871.

[12] Herath S, Harandi M, Porikli F. Learning an invariant hilbert space for domain adaptation [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE,2017:3956 - 3965.

[13] Nguyen T T T, Nguyen T T, Xuan C P, et al. A novel online bayes classifier [C]//2016 International Conference on Digital Image Computing: Techniques and Applications. IEEE,2016:1 - 6.

[14] Dora S, Suresh S, Sundararajan N. Online meta-neuron based learning algorithm for a spiking neural classifier [J]. Information Sciences,2017,414(1):19 - 32.

[15] Seera M, Lim C P, Tan S C. A hybrid FAM-CART model for online data classification [J]. Computational Intelligence, 2018,34(1):562 - 581.

[16] Er M J, Venkatesan R, Wang N. An online universal classifier for binary, multi-class and multi-label classification [C]//2016 IEEE International Conference on Systems, Man and Cybernetics. IEEE,2016:3701 - 3706.

[17] 王炯滔,金明,李有明,等. 基于 Friedman 检验的非参数协作频谱感知方法[J]. 电子与信息学报,2014,36(1):61 - 66.

[18] Garcia S, Fernandez A, Luengo J, et al. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power [J]. Information Sciences, 2010,180(10):2044 - 2064.

[19] Arias J, Gamez J A, Nielsen T D, et al. A pairwise class interaction framework for multilabel classification [M]//Gaag L C, Feelders A J. PGM: European Workshop on Probabilistic Graphical Models. Springer,2014:17 - 32.