文章编号:1672-3961(2011)04-0044-05

# 一种挖掘概念漂移数据流的模糊积分集成分类方法

琚春华[1,2],陈之奇[1*]
(浙江工商大学 1. 计算机与信息工程学院; 2. 现代商贸研究中心, 浙江 杭州 310018)

**摘要**:针对隐含概念漂移和噪声的数据流,提出一种基于模糊积分融合的数据流分类方法(fuzzy integral ensemble classifiers for mining data streams, FI-MDS)。将模糊积分融合方法与集成综合技术有效结合起来,首先通过基分类器对识别样例进行分类得到决策剖面,然后再用模糊积分融合方法得到最终的分类结果,同时引入动态权值更新以提高算法的适应性。实验结果表明,与传统的数据流分类算法相比,该方法提高了概念漂移的检测精度,有效地解决了数据流中复杂分类问题,具有良好的分类性和健壮性。

**关键词**:数据挖掘;数据流;概念漂移;模糊积分
**中图分类号**:TP391 **文献标志码**:A

# A method of fuzzy integral ensemble classifiers for handling concept-drifting data streams

JU Chun-hua[1,2], CHEN Zhi-qi[1*]
(1. School of Computer Science & Information Engineering;
2. Center for Studies of Modern Business, Zhejiang Gongshang University, Hangzhou 310018, China)

**Abstract**: A new classification algorithm FI-MDS based on fuzzy integral fusion was proposed, which aimed at mining data steams with concept drifts and noise and combined fuzzy integral fusion and ensemble multi-classifiers technology. First, the decision-making profile was obtained by training samples through base classifiers, and then the final classification result was obtained via fuzzy integral fusion. Also, a dynamic weight update was introduced to improve the adaptability of this algorithm. Experimental results indicated that this method could enhance the detection accuracy of the concept drifts. Complex classification problems in data streams could be solved and the algorithm has higher classification performance, effectiveness and robustness.

**Key words**: data mining; data streams; concept drift; fuzzy integral

## 0 引言

现实生活中,随着计算机信息技术在商业经济、电子商务、网络安全、证券和市场营销等行业的广泛应用导致了大量数据流的涌现,并以几何递增的方式增长。如:超市交易记录、Web 日志数据、传感器网络、股票交易信息、信用卡交易信息等,这些数据流中蕴含着大量有价值的信息和知识,并且具有海量性、实时性、漂移性等特点。

近年来,流数据分类已成为数据挖掘研究的热点之一,其目标是利用训练数据集建立一个分类预测模型,然后利用该模型对新的数据进行分类预测。由于数据流随着时间的持续变化,这种数据特性的改变使得目标分类模型随时间而改变,进而引起概念漂移(concept drift)问题。传统的静态数据分类

**收稿日期**:2011-02-14
**基金项目**:国家自然科学基金资助项目(71071141);浙江省自然科学基金重点资助项目(Z1091224);浙江省教育厅资助项目(Y201016434)
**作者简介**:琚春华(1962 - ),男,浙江常山人,教授,博士,博士生导师,主要研究方向为人工智能、智能信息处理、电子商务. E-mail:jch@mail.zjgsu.edu.cn
*通讯作者:陈之奇(1984 - ),男,浙江杭州人,硕士研究生,主要研究方向为信息系统与智能信息处理. E-mail:zhiqich@163.com

方法如决策树、决策规则、关联分类法、SVM(support vector machine)等都已经无法满足数据流的处理。因此,数据流中的分类问题备受关注,很多学者对此进行了深入的研究。

本研究提出了一种基于模糊积分融合的数据流分类方法 FI-MDS (fuzzy integral ensemble classifiers for mining data streams),利用模糊积分融合技术将各分类器的输出结果进行融合得到最终的分类结果,并通过集成分类器的动态衰减和更新机制来适应数据流中概念漂移的现象。

## 1　现状分析

Widmer G 和 Kubat M 等于 1996 年提出概念漂移[1]的问题,之后在该领域有很多学者对其进行了研究[2-6];2000 年,Domingos P 等提出增量决策树算法 VFDT(very fast decision tree)[7],Hulten G 等在此基础上对 VFDT 算法进行改进,提出基于决策树模型的分类概念漂移发现算法 CVFDT(conceptadaption very fast decision tree)[8];2001 年,Street W N 等提出一种集成分类器算法 SEA(schoof elkies atkin)[9],该算法是基于批处理方式,按数据流到达时间将其划分成等大的有序数据块,对每块数据采用启发式的替换机制来更新分类器。但该算法每次最多只能替换一个基分类器,不能及时检测数据流中快速的概念变化;2003 年,Wang Z 等人提出了基于加权的集成分类器[10]挖掘概念漂移的数据流。该方法根据各基分类器的分类精度使用带权重的投票算法对其进行加权,并从理论上证明在发生概念漂移的情况下,集成分类器比单一分类器性能更好。该算法的问题在于当训练数据不足时各分类器精度不高,进而影响最终的分类结果;2005 年,Kolter 等人提出 AddExp 集成算法[11],引入了新分类器权重因子和权重衰减因子,并能够逐条处理新到来的数据,在已有集成分类器误分类某条数据时及时调整个体分类器的权重,进而进行增量式学习,以提高算法对概念漂移的检测速度。然而,算法的性能受分类器权值因子和权重衰减因子的影响较大;2008 年,孙岳等提出了一种基于动态自适应修改决策权值参数的增量式多分类器算法 M_ID4[12],利用多分类器综合技术,使用尽量少的训练样本,实现在大容量数据流挖掘中快速地检测概念漂移,但在阈值的确定上还值得深入思考。

通过对国内外已有研究理论分析,发现现实生活中的数据流存在信息的不确定、不完整、模糊性,考虑到真实的数据流可能由于外部原因而包含噪声,故提出一种新的数据流上的分类方法 FI-MDS。该方法提高了概念漂移的检测精度,并且对含有噪声的数据流具有更高的分类能力和健壮性,有效地解决了数据流中的复杂分类问题。

## 2　流数据集成分类模型

### 2.1　相关定义

**定义 1**[13-15]　数据流。令 $t$ 表示任一时间戳,$\boldsymbol{x}_t$ 表示在该时间戳到达的数据向量,数据流可以表示为$\{\cdots,x_{t-1},x_t,x_{t+1},\cdots\}$。

**定义 2**[16-17]　概念漂移。设在时刻 $i$ 处和 $i+1$ 处接收到的数据分别为 $x_i$ 和 $x_{i+1}$,$E_i^*(x_i)$ 和 $E_{i+1}^*(x_{i+1})$分别为数据流在时刻 $i$ 处和 $i+1$ 处的最佳分类模型。如果 $E_i^*(x_i)$ 和 $E_{i+1}^*(x_{i+1})$是不一致的,称数据流从时刻 $i$ 到时刻 $i+1$ 存在概念漂移。

**定义 3**[18]　隶属度。若对论域 $U$ 中的任一元素 $x$,都有一个数 $A(x)\in[0,1]$与之对应,则称 $A$ 为 $U$ 上的模糊集,$A(x)$称为 $x$ 对 $A$ 的隶属度。

**定义 4**[19]　模糊测度。对于测度空间,定义$(X,\Omega)$,其中 $X$ 是一个非空集合,$\Omega$ 是由 $X$ 的若干子集组成的非空类,模糊测度是定义在 $\Omega$ 上的一个非负广义实值函数 $\mu,\mu:\Omega\to[0,\infty]$,且满足下面的条件:

(1) $\mu(\varphi)=0$;

(2) $\forall E\in\Omega,F\in\Omega$,如果 $E\subseteq F$,则 $\mu(E)\leqslant\mu(F)$;

(3) $\forall\{E_n\}\in\Omega(n=1,2,\cdots,\infty)$,有 $E_1\subseteq E_2\subseteq\cdots,\bigcup_{n=1}^{\infty}E_n\in\Omega$,那么$\lim_n\mu(E_n)=\mu(\bigcup_{n=1}^{\infty}E_n)$;

(4) $\forall\{E_n\}\in\Omega(n=1,2,\cdots,\infty)$,有 $E_1\supseteq E_2\supseteq\cdots,\bigcup_{n=1}^{\infty}E_n\in\Omega$,那么$\lim_n\mu(E_n)=\mu(\bigcap_{n=1}^{\infty}E_n)$,模糊测度 $\mu(E)$可以看做 $x\in E$ 的程度;当 $\mu(x)=1$,则称 $\mu$ 为正则模糊测度,通常在多分类器融合中应用的都是正则模糊测度。

**定义 5**[19]　$g_\lambda$ 模糊测度是常用的一种模糊测度,它满足如下性质:若 $A,B\subseteq X$ 和 $A\cap B=\phi$,则有 $g_\lambda(A\cup B)=g_\lambda(A)+g_\lambda(B)+\lambda g_\lambda(A)g_\lambda(B)$, $\lambda>-1$。

**定义 6**[19]　Choquet 积分。令 $\mu$ 是定义在 $X$ 上的模糊测度,$f$ 是定义在 $X$ 上的非负实值可测函数,则 $f$ 关于 $\mu$ 的 Choquet 模糊积分 定义为

$$\int f\mathrm{d}\mu=\sum_{i=1}^{n}\{f(x_i)-f(x_{i-1})\}\mu(A_i),$$

其中,$0\leqslant f(x_1)\leqslant\cdots\leqslant f(x_n)\leqslant 1$,$f(x_0)=0$。

### 2.2 算法思想及设计

基于模糊积分融合的流数据集成分类模型,要求各个基分类器对识别的样例进行分类,得到一个非负实数向量(一般输出值在[0,1]区间内),例如对于一个 $n$ 类分类问题,分类器的输出是非负实数值的 $n$ 维向量,即第 $i$ 个分类器的输出如下形式:$\boldsymbol{d}_i = [d_{i1}, d_{i2}, \cdots, d_{in}]$。这里 $d_{ij}$ 的第一个下标 $i$ 表示分类器 $E_i$ 的输出,第二个下标表示分类器 $E_i$ 对识别样例隶属于的各个类 $C_j$,其中 $d_{ij} \in [0,1]$($j = 1,2,\cdots,n$)。然后,将所有的分类器作为一个集合,记作 $E = \{E_1, E_2, \cdots, E_l\}$。可以确定每一类 $C_i$ 在集合 $E$ 的幂集上的模糊测度 $\mu_i$,就可以用模糊积分把各个分类器对识别样例 $x$ 的输出进行综合,得出待识别样例属于各个类的最终可能性 $e_i$。在计算样例属于 $C_i$ 类的隶属度时,可以把样例决策剖面的第 $i$ 列 $[d_{i1}, d_{i2}, \cdots, d_{in}]^{\mathrm{T}}$ 看作集合 $E$ 上的函数 $f_i$,计算 $f_i$ 关于模糊测度 $\mu_i$ 的模糊积分,该算法选取了 Choquet 积分。最后把最大隶属度对应的类作为样例的最终分类结果,$\lambda^* = \arg(\max_{1 \leqslant i \leqslant c}\{e_i\})$。

为了适应具有概念漂移特征的数据流,集成分类模型引入了动态权值更新机制,这里采用了典型的 Hedge $\beta$ 方法。它是通过对错误预测的分类器进行衰减,保持那些预测正确的分类器权值不变的方法。在多分类器的更新和裁剪问题上,该算法模型是增量式更新的。具体的算法模型如图 1 所示。

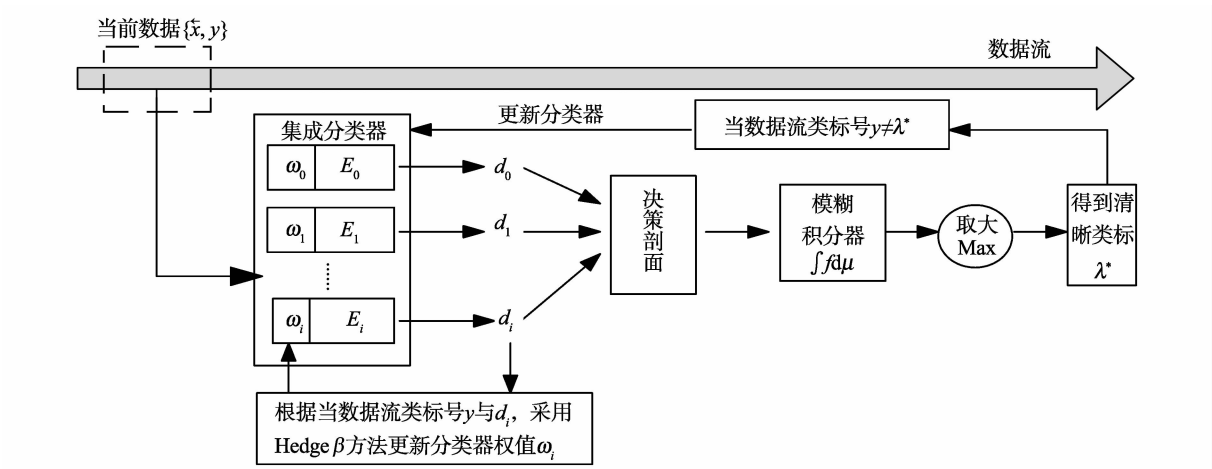


图 1 FI-MDS 算法模型

Fig. 1 FI-MDS algorithm model

### 2.3 FI-MDS 算法描述及步骤

根据 FI-MDS 模型图和基于模糊积分的相关理论,算法描述如下:

**输入** 数据流 $\{\dot{x}, y\}^n$,$y$ 为类标号,

集成分类器 $E = \{E_1, E_2, \cdots, E_l\}$,

当前分类器容量 $m$,

衰减度权重 $\beta, \beta \in [0,1]$,

新分类器更新权值 $\gamma, \gamma \in [0,1]$,

分类器数目阈值 $\theta$。

**输出** 学习后的集成分类器 $E$ 和相应权值 $\omega_i$。

**Step 1** 初始化。

① 初始化变量 $m = 1$;

② 初始化单分类器权值 $\omega_1 = 1$。

**Step 2** IF ($m > \theta$),

选取前 $\theta$ 个大的 $\omega_i$ 对应的 $E_i$ 构建分类器集。

ELSE

选取 $m$ 个 $E_i$ 构建分类器集。

**Step 3** For each $E_i \in E$。

对输入的数据流 $\dot{x}$ 要经过分类器集的每个个体分类器 $E_i$ 进行判别,生成当前数据所对应每个类别的隶属度向量 $\boldsymbol{d}_i = [d_{i1}, d_{i2}, \cdots, d_{in}]$,取 $\boldsymbol{d}_i^* = \arg(\max_{1 \leqslant i \leqslant c}\{\boldsymbol{d}_i\})$。IF($\boldsymbol{d}_i^* \neq y$),$\omega_i = \beta \times \omega_i$。

**Step 4** 由隶属度向量生成决策剖面 DP 以及 $g_\lambda$ 模糊测度。

**Step 5** 利用模糊积分融合多个分类器的结果,由 $\lambda^* = \arg(\max_{1 \leqslant i \leqslant c}\{e_i\})$ 判别 $E$ 的整体决策。

**Step 6** IF($\lambda^* \neq y$),

训练一个新的分类器 $E_{i+1}$,且令 $\omega_{i+1} = \gamma \sum_{i=1}^{n} \omega_i$。转到 Step2。

ELSE 输出结果。

**Step 7** 所有的分类器 $E_i$ 在样本数据流 $\{\dot{x}, y\}^n$ 上增量更新。

## 3 实验分析

本研究的算法使用 Java 编程语言实现,实验环境为 Intel(R) Core(TM)2 Duo CPU E7500 @ 2.93

GHz,2.00 GB 的内存,操作系统是 Windows XP。基分类器训练使用文献[20]的 fVFDT,缓存的大小采用缺省设置。根据数据流具有概念漂移的特点,实验使用了数据流分类算法的经典数据集 SEA[9]。

此数据集具有 3 个属性变量,每个数据项可以看成三维空间中的一个点$(f_1,f_2,f_3)$,$f_i \in \mathbf{R}$,且值范围为 1 ~ 10。如果样本属性满足条件 $f_1+f_2 \leqslant \theta$(阈值 $\theta$ 为一实数),则分配实例类别为 1,否则类别为 0。本实验分别取 $\theta$ 为 8,9,7 和 9.5,来表示 4 个概念、3 次概念漂移。根据每个概念生成 15 000 个实例数据,最后从每个概念块内取 2 500 个实例作为固定的测试集,实验分别在无噪音和含有 5% 噪音两种情况下进行。

### 3.1　性能评测

本实验采用正确率、查准率以及查全率等指标来衡量分类器性能,对分类器的评测采用 $k$-fold ($k=7$)交叉验证方法,其过程如图 2 所示。

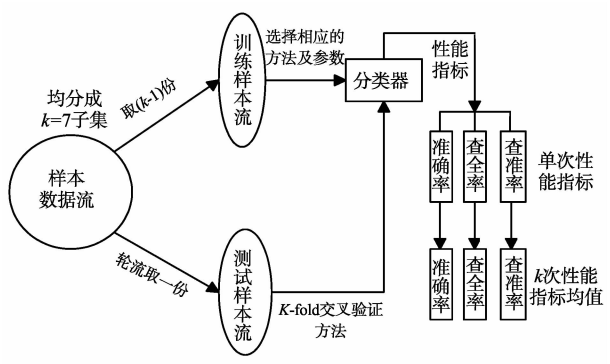


图 2　分类器评测过程

Fig. 2　Evaluating classifiers process

计算方式为

$$P_{\text{accuracy}}=\frac{P_0+N_1}{P+N}, \tag{1}$$

$$P_{\text{precision}}=\frac{P_0}{P_0+N_0}, \tag{2}$$

$$P_{\text{recall}}=\frac{P_0}{P}, \tag{3}$$

其中,假设测试集中包含 $P$ 个正例和 $N$ 个反例,正例中包含被分类器正确分类的 $P_0$ 个样本以及分类错误的 $P_1$ 个样本,反例中包含被分类器误认为正例的 $N_0$ 个样本以及正确识别为反例的 $N_1$ 个样本。

### 3.2　实验结果

分别考察了 FI-MDS 算法与文献[9]的 SEA 算法在无噪音和含 5% 随机噪音情况下的准确率,如图 3 所示,结果表明:由于 FI-MDS 算法采取了模糊积分的融合技术,在无噪音和含有 5% 噪音的情况下,FI-MDS 算法具有更好的准确度和健壮性。

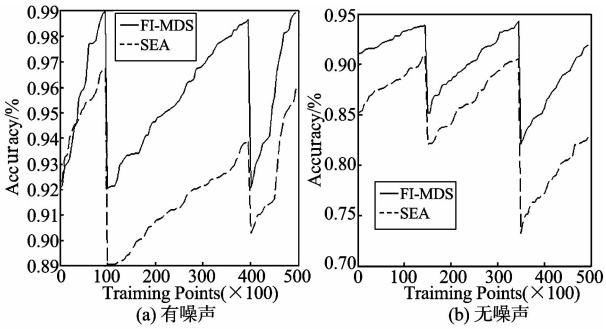


图 3　在无噪声和有噪声情况下的实验结果

Fig. 3　The results under no noise and noise circumstances

对 FI-MDS 算法与 SEA 算法的 $k$-fold 交叉验证实验,分别得到在无噪音和有 5% 噪音环境下的正确率、查准率以及查全率的比较结果,如图 4 所示。

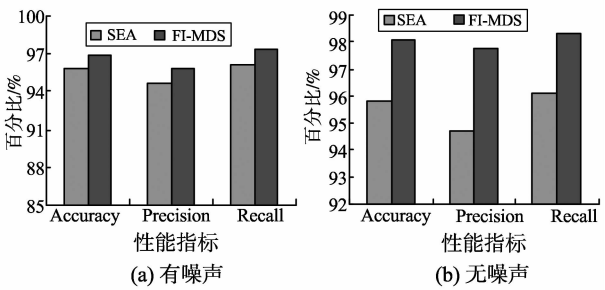


图 4　在无噪声和有噪声情况下的分类性能比较

Fig. 4　The comparison of classification performance under no noise and noise circumstances

## 4　结语

目前多数数据流分类器没有考虑实际生活中数据流信息的不确定、不完整、模糊性,因此本研究提出一种新的数据流上的分类方法——基于模糊积分融合的数据流分类模型 FI-MDS。该方法能够动态适应数据流概念漂移的特性,并在含有噪声数据的环境下具有更好的分类能力,提高了分类精度。本研究在模糊理论的基础上与集成分类器进行有效地结合,但由于算法考虑的不是非常全面,因此还可以进一步得到优化,如在模糊测度的动态更新方面,可以根据待识别样本的不同,模糊测度也随之改变,提高数据流概念漂移的适应性,尚待进一步研究。

**参考文献:**

[1] WIDMER G, KUBAT M. Learning in the presence of concept drift and hidden contexts[J]. Machine Learning, 1996, 23(1): 69-101.

[2] WIDMER G, KUBAT M. Effective learning in dynamic environments by explicit context tracking[C]//Proceeding of the European Conference on Maching Learning. London, UK: Springer-Verlag, 1993:227-243.

[3] WIDYANTORO D H, LOERGER T R, YEN J. An adaptive algorithm for learning changes in user interests [C]//Proceedings of the 8th International Conference on Information and Knowledge Management. Missouri, New York: ACM Press, 1999:405-412.

[4] WIDER G. Tracking context changes through meta-learning[J]. Machine Learning, 1997, 27(3): 259-286.

[5] GANTI V, GEHRKE J, RAMAKRISHNAN R. Mining data streams under block evolution[C]// Proceedings of SIGKDD Explorations. New York:ACM Press, 2002:1-10.

[6] GAMA J, MEDAS P, CASTILLO G, et al. Learning with drift detection[C]// In SBIA Brazilian Symposium on Artificial Intelligence. [S. l.]: Springer, 2004:286-295.

[7] DOMINGOS P, HULTEN G. Mining high-speed data streams[C]//Proceedings of the Association for Computing Machinery Sixth International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2000:71-80.

[8] HULTEN G, SPENCER L, DOMINGOS P. Mining time-changing data streams[C]//Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2001: 97-106.

[9] STREET WN, KIM YS. A streaming ensemble algorithm for large-scale classification [C]//Proceeding of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2001:377-382.

[10] WANG HAIXUN, FAN WEI, YU PHILIP S, et al. Mining concept-drifting data streams using ensemble classifiers[C]//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2003: 226-235.

[11] JEREMY Z KOLTER, MARCUS A MALOOF. Using additive expert ensembles to cope with concept drift [C]//Proceedings of the 22nd International Conference on Machine Leaning. Bonn Germany, New York: ACM Press, 2005:449-456.

[12] 孙岳, 毛国君, 刘旭,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.
SUN Yue, MAO Guojun, LIU Xu, et al. Mining concept drifts from data streams based on multi-calssfifiers [J]. Acta Automatica Sinica, 2008, 34(1): 93-97.

[13] GANTI V, GEHRKE J, RAMAKRISHNAN R. Mining data streams under block evolution[J]. SIGMOD Explorations, 2002, 3(2):1-10.

[14] GUHA S, MEYERSON A, MISHRA N, et al. Clustering data streams: theory and practice[J]. Knowledge and Engineering, IEEE Transactions, 2003, 15(3):515-528.

[15] BABCOCK B, BABU S, DATAR M, et al. Models and issues in data stream systems[C]//Proceeding of PODS. New York, USA: ACM Press, 2002:1-16.

[16] KLINKENBERC R. Learning drifting concepts:example selection vs. example weighting[J]. Intelligent Data Analysis, 2004, 8(3):281-300

[17] FAN Wei. Systematic data selection to mine concept-drift data streams[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM Press, 2004:128-137.

[18] ATANASSOV K T. Intuitionistic fuzzy sets[J]. Fuzzy Sets and Systems, 1986, 20(1):87-96.

[19] WANG Zhenyuan, KLIR George J. Fuzzy measure theory [M]. [S.l.]: Springer, 1992.

[20] 王涛, 李周军, 胡小华,等. 一种高效的数据流挖掘增量模糊决策树分类算法[J]. 计算机学报, 2006, 30(8):1244-1250.
WANG Tao, LI Zhoujun, HU Xiaohua, et al. An incremental fuzzy decision tree classification method for data streams mining based on threaded binary search trees [J]. Chinese Journal of Computers, 2006, 30(8): 1244-1250.

(编辑:孙培芹)