

实验资源篇

本篇主要介绍数据流研究及应用的相关软件平台及数据资源。

第 10 章 数据流分类算法实验工具包 ETDSV1.0.....	314
10.1 引言.....	314
10.2 软件的配置、运行与功能	315
10.2.1 软件的配置与运行	315
10.2.2 软件功能.....	316
10.3 GENERATORS (数据生成器)	318
10.3.1 视图界面中数据生成器主菜单	318
10.3.2 数据库两大生成器菜单功能介绍	319
10.4 SRMTDS 算法	322
10.4.1 SRMTDS 算法参数设定菜单	322
10.4.2 SRMTDS 算法特征数据库读取与算法运行菜单	325
10.5 SRMTCD (MSRT) 算法	327
10.5.1 SRMTCD (MSRT)算法参数设定菜单	327
10.5.2 SRMTCD (MSRT)算法特征数据库读取与算法运行菜单.....	330
10.6 EDT 算法	332
10.6.1 EDT 算法参数设定菜单	332
10.6.2 EDT 算法特征数据库读取与算法运行菜单	336
10.7 EDTC 算法.....	338
10.7.1 EDTC 算法参数设定菜单	339
10.7.2 EDTC 算法特征数据库读取与算法运行菜单	341
10.8 CDRDT 算法.....	344
10.8.1 CDRDT 算法参数设定菜单	344
10.8.2 CDRDT 算法特征数据库读取与算法运行菜单.....	347
10.9 DWCDs 算法.....	349
10.9.1 DWCDs 算法参数设定菜单	349
10.9.2 DWCDs 算法特征数据库读取与算法运行菜单.....	352
10.10 布局图与流程图	353
10.10.1 数据流实验工具算法布局图	353
10.10.2 数据流分类算法流程图	354
第 11 章 经典的数据流分类算法实验工具	355
11.1 VFML 系统	355
11.1.1 VFDTc 算法	355
11.1.2 CVFDT 算法.....	360
11.2 MOA	364
11.2.1 MOA 的界面操作.....	365
11.2.2 MOA 命令行使用方法.....	378
参考文献.....	379
第 12 章 数据流分类算法常用的实验数据集	381
12.1 非概念漂移数据流.....	381
12.1.1 合成数据集.....	381

12.1.2 真实数据集.....	381
12.2 概念漂移数据集.....	382
12.2.1 合成数据集.....	382
12.2.2 真实数据集.....	384
参考文献.....	387

第 10 章 数据流分类算法实验工具包

ETDSv1.0

本章详细介绍了数据流分类算法实验工具包 ETDSv1.0 的功能与用户使用手册。

10.1 引言

数据流分类算法实验工具包 ETDS (an Experimental Tool of Data Stream classification algorithms) (版本号 v1.0.0.0)是由合肥工业大学计算机与信息学院数据挖掘与智能计算“千人计划”DMiC 团队 (HFUT Data Mining and Intelligent Computing Laboratory) 设计开发的一个实验平台,其目的在于为对数据流分类问题研究感兴趣的研究者们提供数据流以及概念漂移数据流分类算法的实验平台与二次开发的平台。

所开发的底层数据结构基于 VFML 的开源代码 (<http://sourceforge.net/projects/vfml/>),包括 DecisionTreeNode (结点类)、ExampleSpec (事例类)等。将此源代码 (C 语言版)部分移植到 VS2005 环境,并将 C 语言风格改称 C++标准语句。在此基础上,开发设计了基于随机决策树模型的数据流及概念漂移数据流 6 大分类算法,包括处理未带有概念漂移问题的数据流算法 SRMTDS 与 EDT 和处理数据流中概念漂移问题的分类算法 SRMTCD、EDTC、CDRDT 与 DWCDs。现将各个算法的特点列举如下:

SRMTDS (Semi-Random Multiple Decision-Tree Algorithm for Data Streams) 算法基于半随机决策树模型,利用 Hoeffding Bounds 不等式及信息熵方法设定连续属性结点的分割阈值,能有效地处理数据流中连续属性问题,同时在叶子结点引入朴素贝叶斯分类器能有效降低分类错误率。该算法能快速适应未带概念漂移的数据流,具有较高的分类正确率与较好的时空性能。

EDT (Ensemble Decision Trees for data streams) 算法的提出与 SRMTDS 算法相似,也是为了处理未带概念漂移数据流的分类问题,而不同在于该算法探索了三种不同的连续属性分割阈值选择方法 (包含 SRMTDS 算法的连续属性分割阈值的选择方法),同时算法中构建的决策树不是预先完成而是伴随着新数据块的到来增量式构建。实验验证了三种不同的连续属性的分割阈值选择方法中基于完全随机的方法具有更佳的时空性能,优于 SRMTDS 算法的分类效果。

SRMTCD (Semi-Random Multiple Decision-Tree Algorithm for Concept-Drifting Data Streams) (又称 MSRT) 算法在 SRMTDS 算法的基础上,考虑数据流具有概念漂移的特点,采用双窗口机制周期性地检测数据流是否发生了概念漂移,同时利用 Hoeffding 边界不等式设定临界阈值区分概念漂移与噪音,调整双窗口的大小以适应具有不同概念漂移特征的数据流。该算法在时间性能、抗噪能力和分类正确率等方面具有一定优势。

EDTC (Ensemble Decision Trees for Concept Drifting data streams) 算法是在 EDT 算法的基础上添加概念漂移检测机制,以使 EDT 算法扩展成为可以适宜具有概念漂移的数据流分

类算法，其概念漂移检测策略基于双阈值机制，目的在于有效跟踪不同类型的概念漂移同时降低噪音对概念漂移检测的影响。实验也验证了该算法对概念漂移数据流处理的有效性。

CDRDT (a streaming data algorithm for Concept Drifts in Random Decision Trees) 算法的提出目的在于进一步提高 SRMTCD 算法与 EDTC 算法对概念漂移数据流处理的时空性能与抗噪性。CDRDT 采用非固定大小的数据块构建集成随机决策树模型，基于 Hoeffding Bounds 不等式与统计质量控制原理设定不同的漂移检测阈值以有效从噪音中区分不同类型的概念漂移。

考虑到数据流中概念漂移类型的多样性，为有效检测出不同类型的概念漂移，进一步探索不同于 SRMTCD、EDTC 与 CDRDT 的概念漂移检测机制，于是 DWCDs (a Double-Window-based Classification algorithm for concept drifting Data Streams) 算法得以诞生。该算法基于 CDRDT 中的完全随机决策树模型，依据窗口中原始数据分布变化来检测概念漂移。

这六大算法均采用集成随机决策树模型作为基本分类器，能有效适应数据流分类及数据流概念漂移检测问题。数据流分类算法实验工具包 ETDS 的开发平台是 VS2005+MFC+WindowsXp，生成了 ETDSv1.0.0.0.exe 可执行文件，用户可通过此界面环境实现界面的交互。

以下章节具体安排如下：10.2 节简要介绍软件的配置、运行与功能；10.3 节相关数据生成器生成界面使用方法；10.4 节简介 SRMTDS 算法运行的用户界面使用方法；10.5 节简介 SRMTCD (MSRT)算法运行的用户界面使用方法；10.6 节简介 EDT 算法运行的用户界面使用方法；10.7 节简介 EDTC 算法运行的用户界面使用方法；10.8 节简介 CDRDT 算法运行的用户界面使用方法；10.9 节简介 DWCDs 算法运行的用户界面使用方法；最后给出了数据流实验工具算法布局图与分类算法流程图。

10.2 软件的配置、运行与功能

本节主要简介软件包 ETDSv1.0.0.0 的用户界面使用环境配置、运行条件与软件包的功能。

10.2.1 软件的配置与运行

(1) 软件的配置

软件配置：

- 1) Windows 9x\me\NT\2000\XP 操作系统；

硬件配置：

- 2) CUP 运行速度 1.0G 以上均可，内存要求不高：64 兆，128 兆以上

(2) 软件的运行

一般在上述软硬件环境下,双击 ETDSv1.0.0.0.exe 即可进入用户交互界面。如出现异常,则需要以下操作:由于此软件包在 VS2005+MFC 环境下开发,可能需要一些动态链接库 mfc70d.dll、mfc71d.dll、msvcr70d.dll、msvcr71d.dll、msvci70d.dll 见图 10.1,将其拷本到可执行文件的同一目录下即可。



图 10.1 所需的动态链接库

待以上步骤完成后,就可以直接双击 ETDSv1.0.0.0.exe,从而进入如图 10.2 所示的界面环境中去。

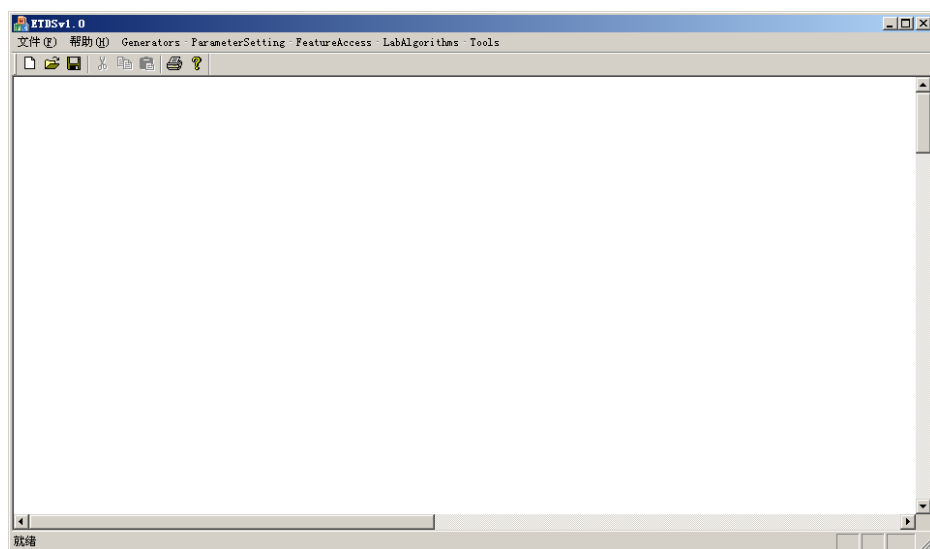


图 10.2 ETDSv1.0 用户界面

10.2.2 软件功能

整个软件包 ETDSv1.0 展示了数据流分类六大算法运行的用户交互界面。由图 10.2 可知:整个可视化实验环境分五大部分:Generators (数据生成器), ParameterSetting (算法参数设定), FeatureAccess (特征信息获取), LabAlgorithms (算法集合) 与 Tools (工具集)。

Generators (数据生成器): 提供两组漂移数据生成器-HyperPlane 和 STAGGER;

ParameterSetting (算法参数设定): 提供六大数据流分类算法--SRMTDS 算法、SRMTCD (MSRT) 算法、EDT 算法、EDTC 算法、CDRDT 算法与 DWCDS 算法的参数设置菜单,如图 10.3 所示;

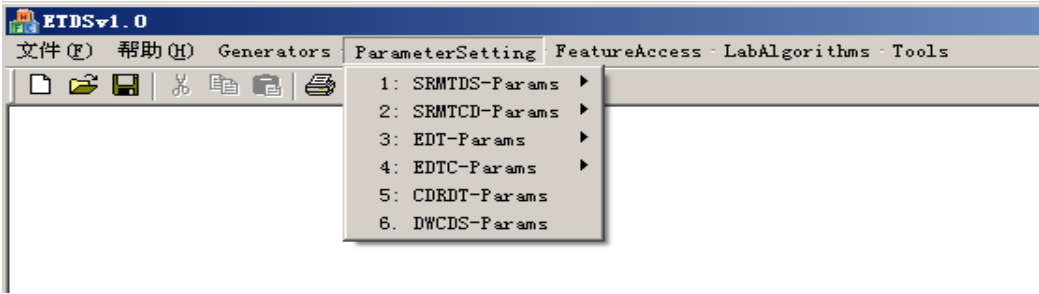


图 10.3 ETDSv1.0 包含 6 大分类算法参数设置菜单功能项

FeatureAccess (特征信息获取): 读取数据库特征文件的菜单如图 10.4 所示;

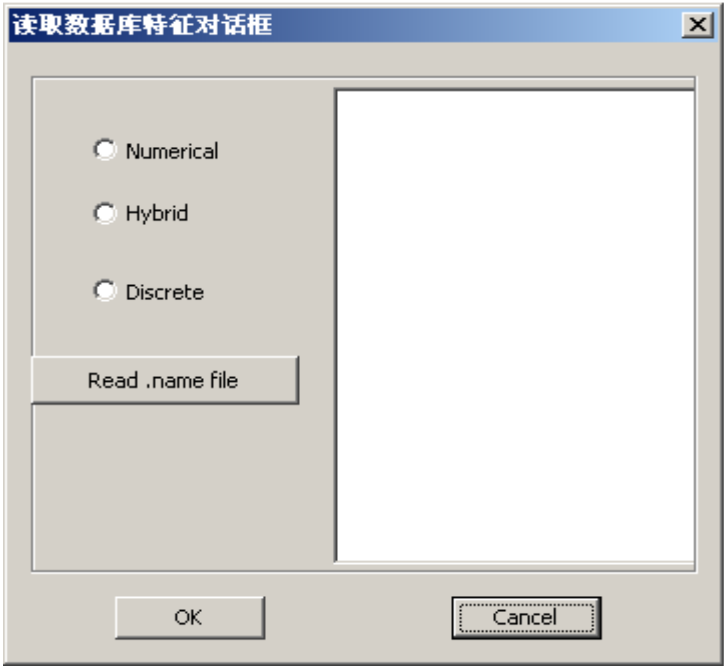


图 10.4 读取数据库特征文件界面

LabAlgorithms(算法集合): 提供数据流分类六大算法-SRMTDS 算法, SRMTCD(MSRT) 算法、EDT 算法、EDTC 算法、CDRDT 算法与 DWCDS 算法的运行交互界面, 如图 10.5 所示;

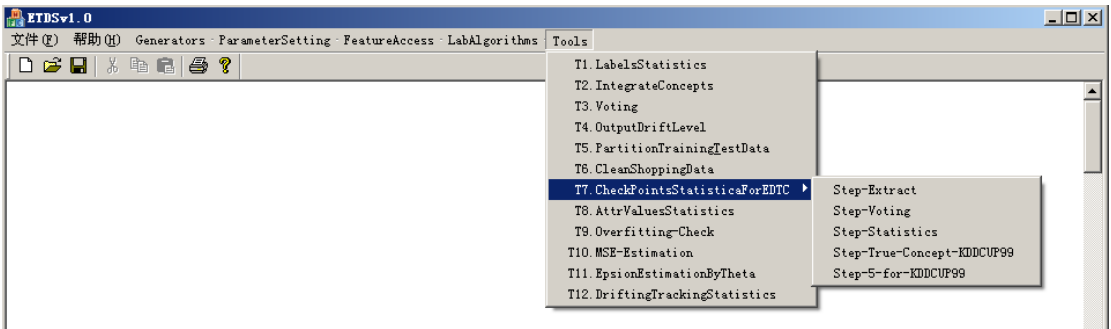


图 10.5 ETDSv1.0 包含的辅助工具菜单功能项

Tools (工具集): 辅助一些函数功能菜单, 包括数据库类别分布统计 (LabelsStatistics)、概念合并菜单 (IntegrateConcepts), 投票 (voting), 输出漂移情况 (OutputDriftLevel) 等功能项, 如图 10.6 所示。

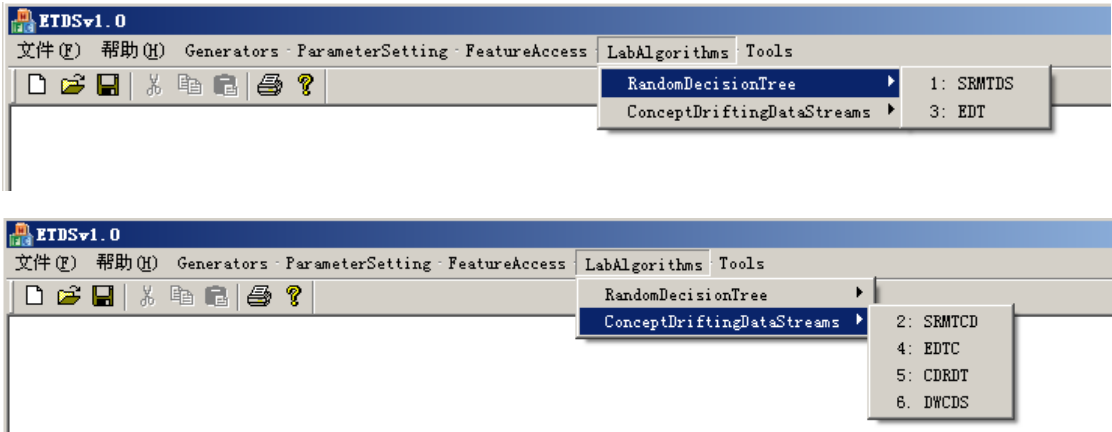


图 10.6 ETDSv1.0 包含 6 大分类算法菜单功能项

10.3 Generators (数据生成器)

本节主要简介软件包 ETDSv1.0.0.0 开发的若干数据生成器的用户界面使用方法。

10.3.1 视图界面中数据生成器主菜单

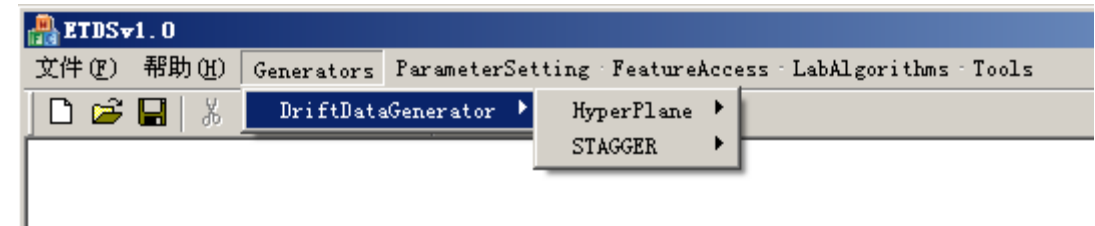
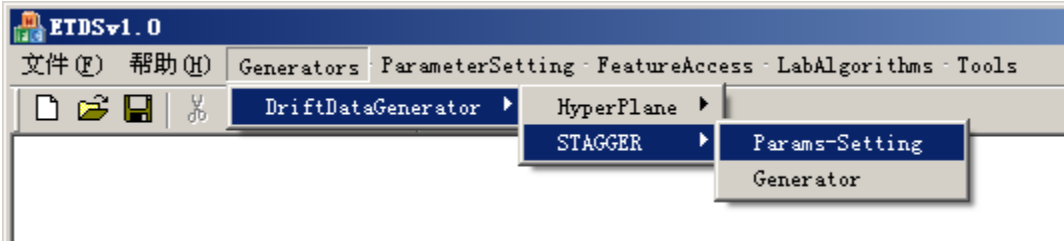


图 10.7 数据生成器菜单功能项

由图 10.7 可知, 数据生成器的菜单项主要包含两大漂移数据库-HyperPlane 与 STAGGER 菜单。使用方法: 鼠标选中四种任一个数据库如: HyperPlane, 出现两个菜单项---参数设定 (Params—Setting)与数据生成(Generator), 如图 10.8 所示。



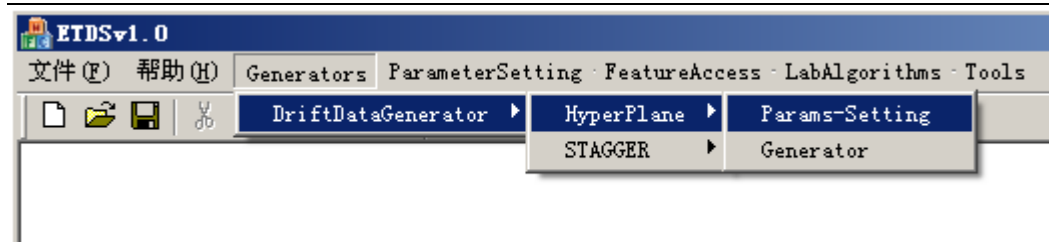


图 10.8 漂移数据库生成器菜单功能项

10.3.2 数据库两大生成器菜单功能介绍

任一数据库的生成均包含两大步骤，即参数设定与数据生成。以下将逐一介绍此两大步骤。

10.3.2.1 参数设置菜单—(Params—Setting)

HyperPlane 数据生成器

关于 HyperPlane 数据库生成器的描述见图 10.9。

选择图 10.8 所示的 HyperPlane 数据库生成器的 Params—Setting 菜单项，弹出如图 10.10 的参数设置对话框，现针对对话框中的每一个功能键对应的参数列表如下表 10.1：

HyperPlane is a benchmark database of data streams with the gradual concept drift. A HyperPlane in a d dimensional space (e.g., $d = 10$) is denoted by equation: $\sum_{i=1}^d w_i x_i = w_0$. Each vector of variables (x_1, x_2, \dots, x_d) in this database is a randomly generated instance and is uniformly distributed in the multidimensional space $[0, 1]^d$. If $\sum_{i=1}^d w_i x_i \geq w_0$, the class label is 1, or else it is 0. The bound of coefficient w_i is limited to $[\min, \max]$. For a weight of w_i till it is up or down to the boundary, then changes the direction with the probability of p_w . In this data set, the noise rate is introduced by $r\%$ and the changed dimensions of attributes is set to t (e.g., $t = 2$).

图 10.9 HyperPlane 数据库生成器描述

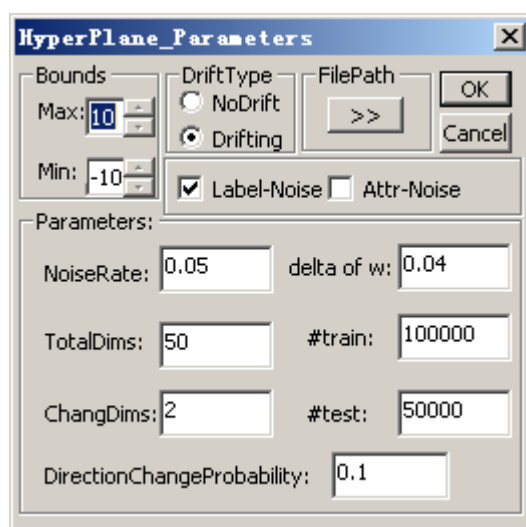


图 10.10 HyperPlane 数据库生成器参数设置对话框

表 10.1 HyperPlane 数据库生成器参数描述

参 数 名称	参数意义	类型	界面对应位置	参数值 (初始值)
~	The path setting of data set	string	按钮 FilePath>>	默认在系统自定义的此运行环境的根目录下, 自定义最佳
<i>max</i>	The largest bound of weight change related to w	int	Max	10
<i>min</i>	The lowest bound of weight change related to w	int	Min	-10
<i>r</i>	Noise rate	double	NoiseRate	0.05 (5%)
<i>d</i>	Total dimensions of attributes	int	TotalDims	50
<i>t</i>	The dimensions of attributes with changing weight	int	ChangdDims	2
Δw	Change rate of weight- w	double	delta of w	0.04
~	The number of training instances	int	#train	100000
~	The number of testing instances	int	#test	50000
<i>pw</i>	The probability of changing direction	double	DirectionChange-Probability	0.1 (10%)
~	Whether the current data set has drifts or not	bool	NoDrift Drifting	默认值: Drifting
~	The type of noise: noise at the class labels or at the dimensions of attributes	bool	Label-Noise Attr-Noise	默认值: Label-Noise

STAGGER 数据生成器

关于 STAGGER 数据库生成器的描述见图 10.11。

STAGGER is a standard database of concept-shifting data streams to test the abilities of inductive algorithms. This database consists of three attribute values: color $\in \{\text{green, blue, red}\}$, shape $\in \{\text{triangle, circle, rectangle}\}$ and size $\in \{\text{small, medium, large}\}$. There are three alternative underlying concepts, A: if color=red \wedge size=small, class=1; otherwise, class=0; B: if color=green \vee shape=circle, class=1; otherwise, class=0; and C: if size=medium \vee large, class=1; otherwise, class=0. The data set generated randomly in our experiments contains 0.1k concepts and each concept contains 1k-sized random instances. The initial concept begins with A, and these three concepts can transfer to each other. The drifting details are as follows: $A \rightarrow B$, $B \rightarrow A$ and $C \rightarrow B$ with a shifting probability of 25%; $A \rightarrow C$, $C \rightarrow A$ with a shifting probability of 75%.

图 10.11 STAGGER 数据库生成器描述

选择图 10.8 所示的 STAGGER 数据库生成器的 Params—Setting 菜单项弹出 STAGGER 参数设置对话框, 如图 10.12 所示, 其参数列表见表 10.2。

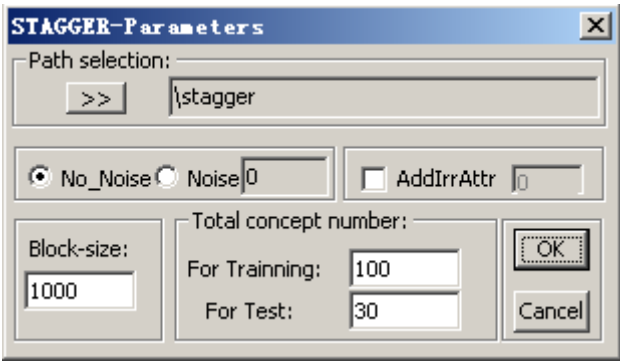


图 10.12 STAGGER 数据库生成器参数设置对话框

表 10.2 STAGGER 数据库生成器参数描述

参数名称	参数意义	类型	界面对应位置	参数值（初始值）
~	设置生成数据库的路径	string	按钮>>	默认在系统自定义的 stagger 文件夹下
min	添加不相关属性维	int	选择框 AddIrrAttr	默认为不添加不相关属性， 即文本框中值为 0
	数据块大小	int	Block-size	1000
~	The number of chunks of training instances	int	For Training	100
~	The number of chunks testing instances	int	For test	30
~	Whether the current data set has label noise or not	bool	No_Noise Noise	默认值：No_Noise

注：利用此数据生成器生成的 STAGGER 数据内容包括：100*1000 大小的训练集，即 100 块数据块，每一块 1000 大小代表一个概念，下一块（概念）的变化规律见图 10.11 的描述。
30*1000 大小的测试集，此测试集包含的概念与训练集相同，变化规律也相同。

10.3.2.2 数据生成菜单—Generator

完成相应的数据库参数设定后，选择图 10.8 所示的对应数据库生成器的 Generator 菜单项，进入生成数据阶段，此过程无界面显示，生成数据完毕后，弹出对话框提示数据生成完毕，如图 10.13。

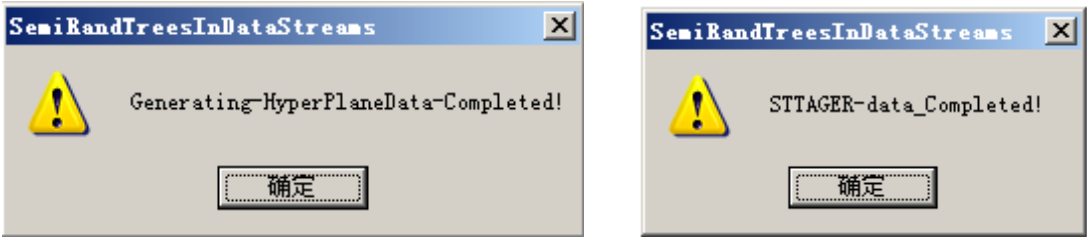


图 10.13 数据库生成完毕对话框

执行 STAGGER 与 HyperPlane 数据生成器生成数据时附带的文件列表见表 10.3。

表 10.3 生成数据库运行后产生的数据文件列表

STAGGER	HyperPlane	含义
STAGGER.concept	HyperPlane.Concept	记录数据生成器参数设置情况的文件
STAGGER.data	HyperPlaneTrain.data	训练数据文件
STAGGER.names	HyperPlane.names	数据库特征文件
STAGGER.test	HyperPlane.test	测试数据文件
	HyperPlaneTest.Noise	测试数据中噪音数据记录文件
	HyperPlaneTrain.noise	测试数据中噪音数据记录文件
	HyperPlane.tmp	生成数据过程中的临时文件，针对 HyperPlane
	true-drifting-points.txt	记录真实漂移位置的文件

10.4 SRMTDS 算法

SRMTDS 算法的运行分为三大步骤，步骤 1：相关参数的设置；步骤 2：数据库特征文件的读取；步骤 3：算法运行。以下将针对界面操作逐步介绍算法运行的三大步骤。

10.4.1 SRMTDS 算法参数设定菜单

由图 10.14 可知，SRMTDS 算法中参数设定按照分割方法不同分为 Hoeffding 方式与

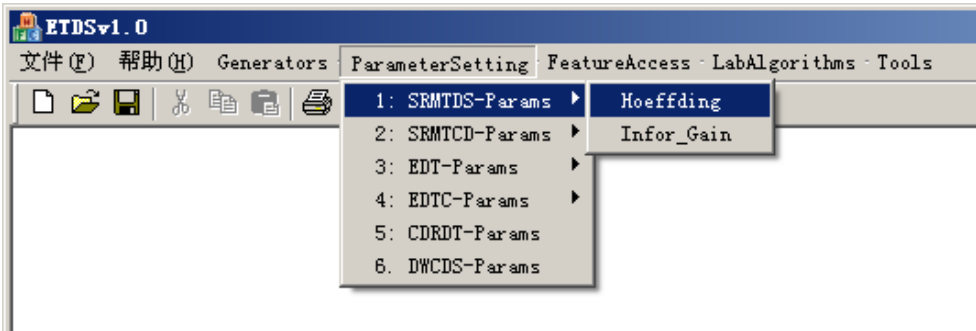


图 10.14 SRMTDS 算法参数设定主菜单

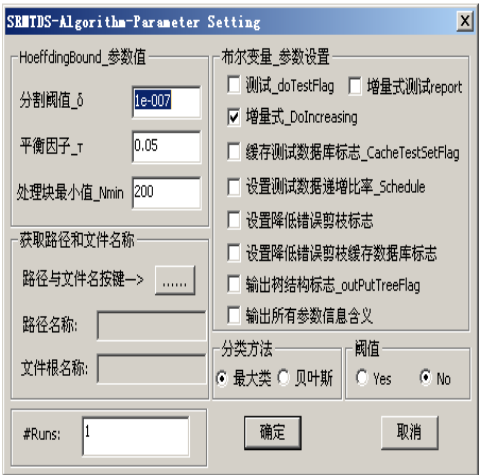


图 10.15 选择 Hoeffding 菜单弹出的对话框

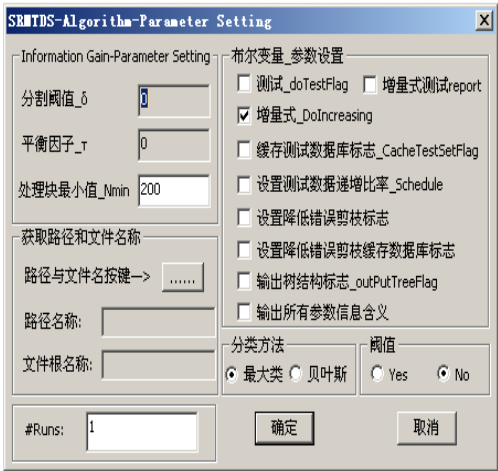


图 10.16 选择 Infor_Gain 菜单弹出的对话框

Infor_Gain 两种方式，选择两种方式的菜单弹出相应的参数对话框，不同在于后者对应的参

数设定中没有 Hoeffding bounds 不等式中的一些参数，如： δ 与 τ 。

现将参数列举如下，见表 10.4。

表 10.4 SRMTDS 算法一般参数描述（针对图 10.15 与图 10.16）

参数名称	参数意义	类型	界面对应位置	参数值（初始值）
m_fSplitConfidence	Hoeffding Bounds 不等式中的分割阈值	double	分割阈值_ δ	10^{-7}
m_fTieConfidence	避免 ties 的阈值	double	平衡因子_ τ	0.05
m_iMinChunk	决策树结点中运行一次分割所需的数据量大小	int	处理块最小值_Nmin	200
~	设置生成数据库的路径，读取.test 文件	string	路径与文件名按键 ->	无默认值
m_ShowSourceDirectoryInEdit	只读项，显示读取的文件名称路径	string	路径名称	
m_ShowFileStemInEdit	只读项，显示读取的.test 文件名称的无后缀名称，用来指示.data 文件名与此文件根名一致	string	文件根名称	
m_iRuns	指示当前算法运行次数	int	#Runs	1
m_bDoIncrementalReport	增量式输出分类结果标志	bool	选择框增量式测试 report	false，即不增量式输出训练中的分类结果
m_bDoIncre	指示当前算法是批量式处理算法还是增量式处理算法	bool	增量式_DoIncreasing	true; true 表示为增量式算法，false 表示为批处理算法
m_bDoTestFlag	/	bool	选择框测试 doTestFlag	false
m_bCacheTestSet	/	bool	缓存测试数据库标志 _CacheTestSetFlag	false，即不设置缓冲测试区标志
m_bUseSchedule	/	bool	设置测试数据递增比率_Schedule	false
m_bREPrune	/	bool	设置降低错误剪枝标志	false
m_bCachePruneSet	/	bool	设置降低错误剪枝缓存数据库标志	false
m_bOutPutTrees	/	bool	输出树结构标志 m_outPutTreeFlag	false
m_bOutPutHelp	/	bool	输出所有参数信息含义	false
m_bBayesMethod	分类方法	bool	最大类	false，即采用最

			贝叶斯	大类方法
m_bConValuesThres	阈值	bool	Yes	false，在选择采用贝叶斯分类方法时，此参数才设置为有效状态
			No	

注：m_bCacheTestSet，m_bUseSchedule，m_bREPrune，m_bCachePruneSet，m_bOutPutTrees，m_bOutPutHelp 等变量主要针对经典 VFDTc 算法，在此处没有特别用处。m_bDoIncrementalReport，m_bConValuesThres 值为 true 时，即处于有效状态时，会相应弹出图 10.17 与图 10.18 的对话框，相应设置增量式输出的事例间隔参数 m_increReportCount（默认值为 10000）与连续属性处理时累积的连续属性值个数阈值 m_maxNumOfConValues（默认值为 1000）。

假设 SRMTDS 算法此时采用 Hoeffding Bounds 不等式作为决策树增长结点分割方法见

图 10.15 界面，在此界面中，选择 **路径与文件名按钮**→ **.....** 按钮，打开***.test 文件（以 adult.test）为例，故以下的操作均是在以选择 adult.test 文件之后的界面交互操作。

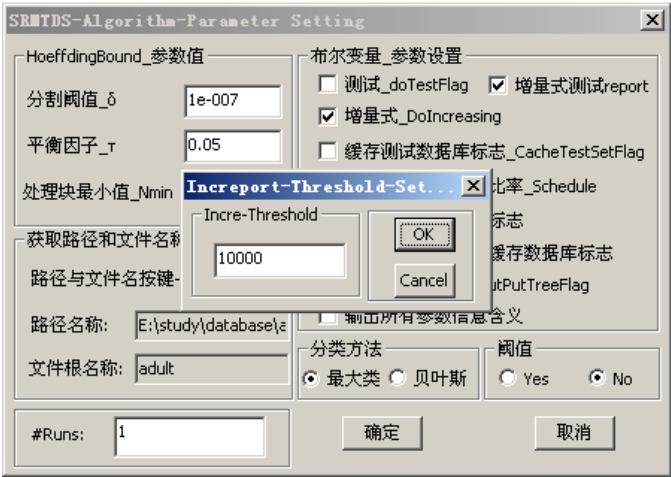


图 10.17 选择增量式测试 report 弹出的对话框

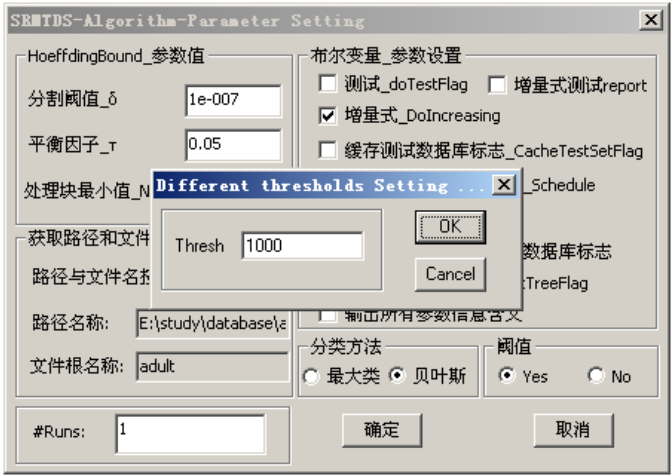


图 10.18 选择贝叶斯分类方法与阈值设定后弹出的对话框



图 10.19 选择图 10.18 中确定按钮后弹出的 SRMTDS 特有参数设定对话框

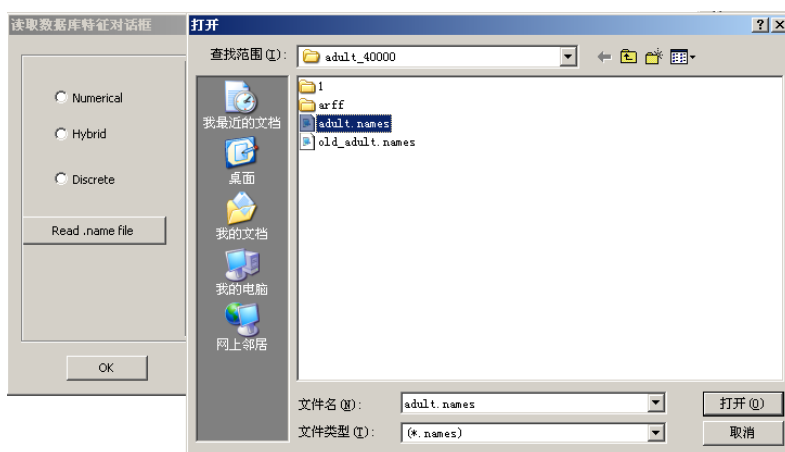
表 10.5 SRMTDS 算法参数描述（针对图 10.19）

参数名称	参数意义	类型	界面对应位置	参数值（初始值）
m_iTreeNum	集成分类器中决策树构建的棵树	int	Tree-Num	5
m_initHeightOfTree	决策树的树高阈值	int	h0	5
/	输出信息阈值	int	Mess-Level	2
/	内存检测阈值	int	IniMax-Alloc	
m_iSplitRandomType	采用半随机策略分割阈值标志	enum	RHB	HOEFFDINGSELECTION, 即采用半随机策略分割阈值
m_iRescanPeriod	重新扫描周期	int	Rescan	500000
m_maxCoefficient	概念漂移检测中系数最大值	int	Max	2
m_minCoefficient	概念漂移检测中系数最小值	int	Min	1
	确定此对话框参数设置有效的按钮		OK 按键	

10.4.2 SRMTDS 算法特征数据库读取与算法运行菜单

以上涉及的参数设置是算法运行的第一步骤，完成参数设置之后，需要进行以下两步骤得以运行相应的算法，即步骤二：读取待处理的数据库的特征（读取 adult.names 文件），见图 10.20-(a)，图 10.20-(b)；步骤三：完成参数设置与特征读取后，选择对应的算法菜单如 SRMTDS 算法，进入选择读取.data 文件的界面，见图 10.20-(c)与图 10.20-(d)，以选择 adult.data 为例，读取了训练集后，选择“打开”按钮，即进入运行 SRMTDS 算法阶段。在算法运行阶段，界面无任何提示状态，一旦算法运行完毕，相应的分类结果在界面显示，如图 10.20-(e)。前 5 个 Test_Result 对应的行表示 SRMTDS 算法中集成分类器各自的分类效果，最后一个

Test_Result 对应的行是算法中构建的集成分类器的最终投票结果，即算法的分类效果。



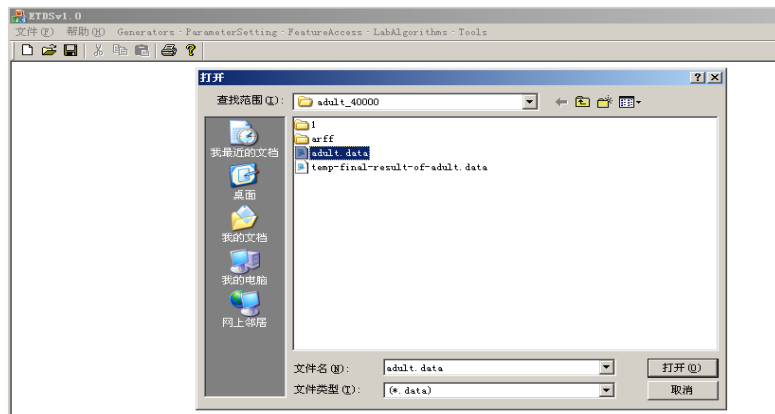
(a) 步骤 2-1—选择 Read.names file 按钮，弹出右侧文件打开框



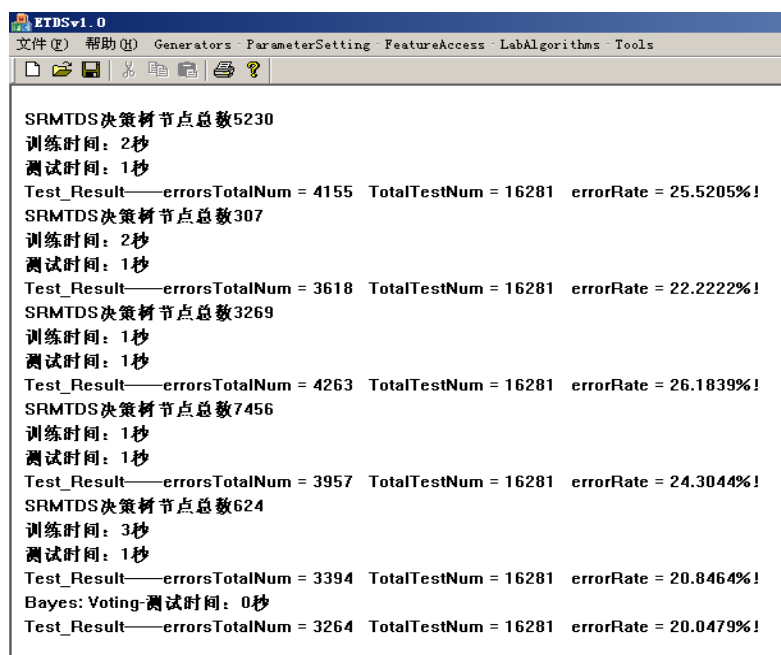
(b) 步骤 2-2—选择步骤 2-1 右侧打开窗口的打开按钮，读取数据库特征并显示



(c) 步骤 3-1—选择 SRMTCD 算法，弹出打开对话框（如下图）



(d) 步骤 3-2—打开训练数据集 adult.data



(e) SRMTDS 算法运行完毕时的分类结果（以 adult 数据库为例）

图 10.20 SRMTDS 算法运行中的特征读取、算法运行以及结果显示

10.5 SRMTCD（MSRT）算法

与 SRMTDS 算法运行步骤相似，SRMTCD（MSRT）算法的运行也分为三大步骤，步骤 1：相关参数的设置；步骤 2：数据库特征文件的读取；步骤 3：算法运行。以下将针对界面操作逐步介绍算法运行的三大步骤。

10.5.1 SRMTCD (MSRT)算法参数设定菜单

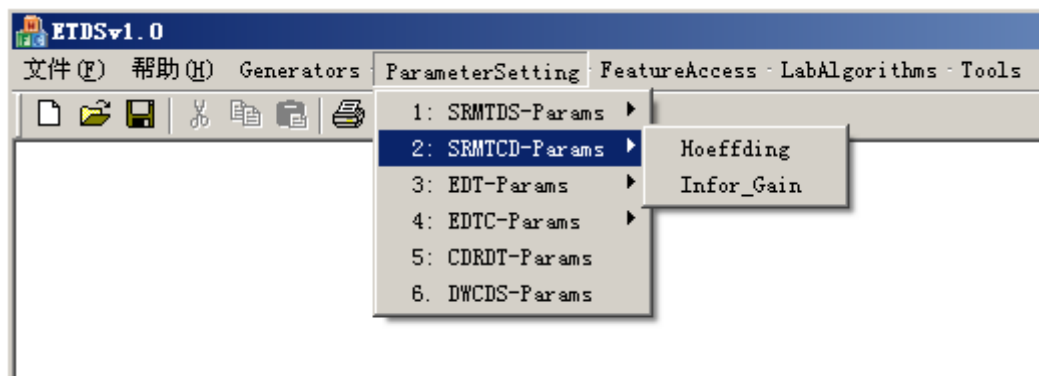


图 10.21 SRMTCD(MSRT)算法参数设定主菜单

由图 10.21 可知，SRMTCD(MSRT)算法中参数设定按照分割方法不同分为 Hoeffding 方式与 Infor_Gain 两种方式，选择两种方式的菜单弹出相应的参数对话框，不同在于后者对应的参数设定中没有 Hoeffding bounds 不等式中的一些参数，如： δ 与 τ ，此种情况与算法 SRMTDS 一致。



图 10.22 选择 Hoeffding 菜单弹出的对话框



图 10.23 选择 Infor_Gain 菜单弹出的对话框

现将参数列举如下，见表 10.6。

表 10.6 SRMTC(MSRT)算法一般参数描述（针对图 10.22 与图 10.23）

参数名称	参数意义	类型	界面对应位置	参数值（初始值）
m_fCVSplitThresh	Hoeffding Bounds 不等式中的分割阈值	double	分割阈值_δ	0.0001
m_fCVTieThresh	避免 ties 的阈值	double	平衡因子_τ	0.05
m_iCVCheckMinNum	决策树结点中运行一次分割所需的数据量大小	int	处理块最小值 _Nmin	300
~	设置生成数据库的路径，读取.test 文件	string	路径与文件名按键 —>	无默认值
m_strCVPathText	只读项，显示读取的文件名称路径	string	路径名称	
m_strCVFileStem	只读项，显示读取的.test 文件名称的无后缀名称，用来指示.data 文件名与此文件根名一致	string	文件根名称	
m_bBayesMethod	分类方法	bool	最大类 贝叶斯	false，即最大类方法
m_iCVCheckSize	检测替换子树是否更适宜当前窗口数据时的检测周期	int	测试例子大小初始值 CheckSize	10000
m_iCVWindowSize	滑动窗口大小	int	窗口初始值 WindowSize	50000
m_iCVCaheSize	/	int	缓存区大小初始值 CacheSize	10000
m_iCVAltTestNum	判断替换子树是否更适宜当前窗口数据时需要的数据集大小阈值	int	替换分支测试数初始值 AltTestNum	1000

m_fCVSheduleMulti	/	double	测试数据测试比例 初始值 SchedulMult	1.44
m_iCVGrowMessg	/	int	总空间消耗初始值 GrowMessage	1000
m_iCVShedulCount		int	测试初始值 SheduleCount	10000
m_CVDoTest	/		doTest	
m_CVDoIncrReport	增量式输出分类结果标志	bool	IncrReport	false; true 表示为 增量式算法, false 表示为批处理算法
m_bCVfinalOutput	/		finOutput	false;
m_iRuns	指示当前算法运行次数	int	#Runs	1

注：m_iCVCaheSize, m_fCVSheduleMulti, m_iCVGrowMessg, m_iCVShedulCount, m_CVDoTest 等变量主要针对经典 CVFDT 算法设定，在此处没有特别用处。m_CVDoIncrReport 值为 true 时，即处于有效状态时，会相应弹出图 10.24 的对话框，相应设置增量式输出的事例间隔参数 m_incrReportCount（默认值为 10000）。

假设 SRMTCD 算法此时采用 Hoeffding Bounds 不等式作为决策树增长结点分割方法见图 10.22 界面，在此界面中，选择 路径与文件名按键—> 按钮，打开***.test 文件（以 adult.test）为例，故以下的操作均是在以选择 adult.test 文件之后的界面交互操作。

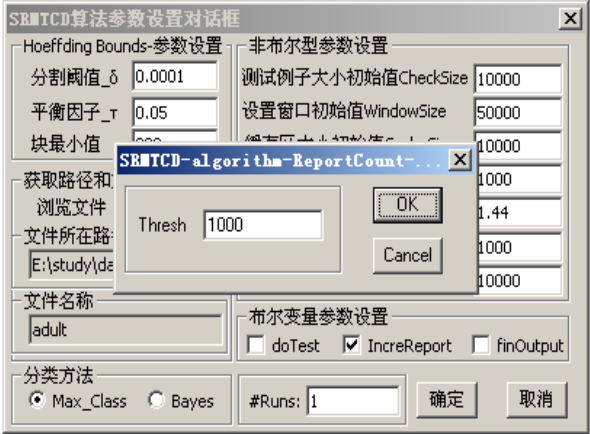


图 10.24 勾选 IncrReport 选择框时弹出的对话框

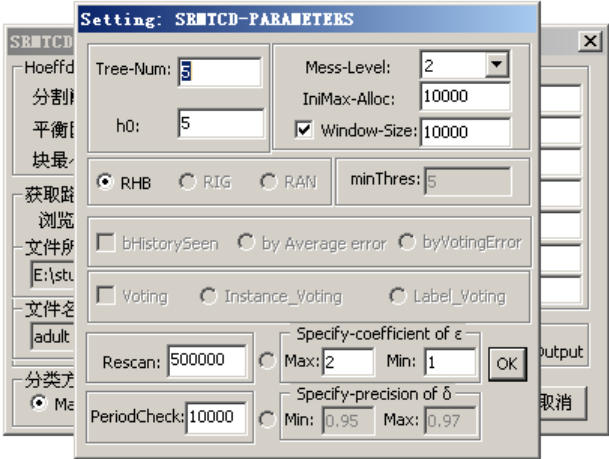


图 10.25 选择图 10.24 中确定按钮后弹出的 SRMTCD 特有参数设定对话框

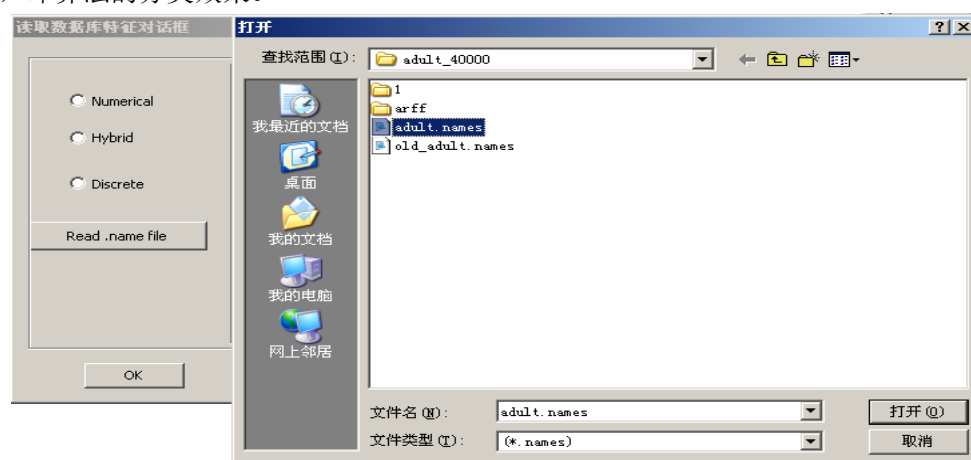
图 10.25 中对话框中参数列表如下，见表 10.7。

表 10.7 SRMTCD(MSRT)算法参数描述（针对图 10.25）

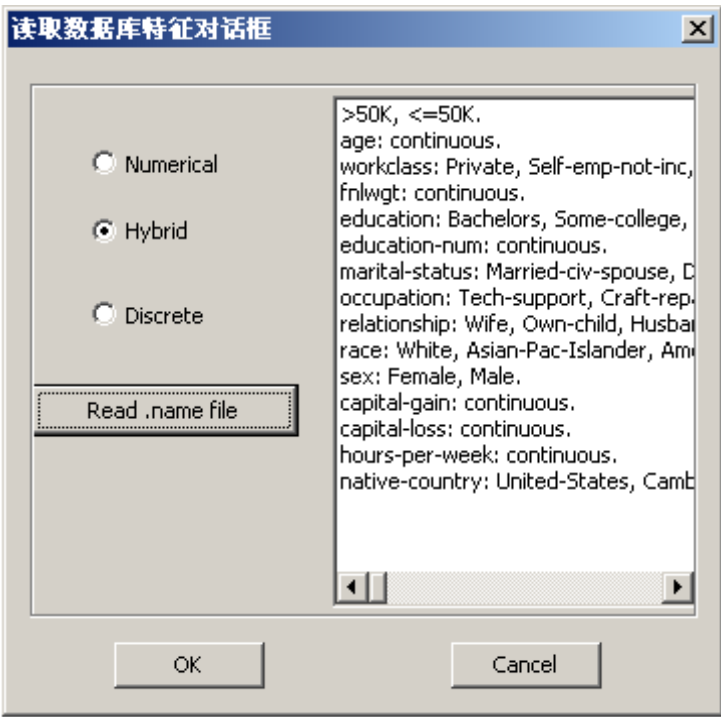
参数名称	参数意义	类型	界面对应位置	参数值（初始值）
m_iTreeNum	集成分类器中决策树构建的棵树	int	Tree-Num	5
m_initHeightOfTree	决策树的树高阈值	int	h0	5
/	输出信息阈值	int	Mess-Level	2
/	内存检测阈值	int	IniMax-Alloc	
m_iSplitRandomType	采用半随机策略分割阈值标志	enum	RHB	HOEFFDINGS ELECTION
m_iRescanPeriod	重新扫描周期	int	Rescan	500000
m_iCheckPeriod	概念漂移检测周期	int	PeroidCheck	10000
m_maxCoefficient	概念漂移检测中系数最大值	int	Max	2
m_minCoefficient	概念漂移检测中系数最小值	int	Min	1
	确定此对话框参数设置有效的按钮		OK 按键	

10.5.2 SRMTCD (MSRT)算法特征数据库读取与算法运行菜单

以上涉及的参数设置是算法运行的第一步骤，完成参数设置之后，需要进行以下两步骤得以运行相应的算法，即步骤二：读取待处理的数据库的特征（读取 adult.names 文件），见图 10.26-(a)，图 10.26-(b)；步骤三：完成参数设置与特征读取后，选择对应的算法菜单如 SRMTDS 算法，进入选择读取.data 文件的界面，见图 10.26-(c) 与图 10.26-(d)，以选择 adult.data 为例，读取了训练集后，选择“打开”按钮，即进入运行 SRMTDS 算法阶段。在算法运行阶段，界面无任何提示状态，一旦算法运行完毕，相应的分类结果在界面显示，如图 10.26-(e)。前面 5 个 Test_Result 对应的行表示 SRMTCD 算法中构建的集成分类器各自的分类效果，最后一个 Test_Result 对应的行是 SRMTCD 算法中构建的集成分类器的最终投票结果，即算法的分类效果。



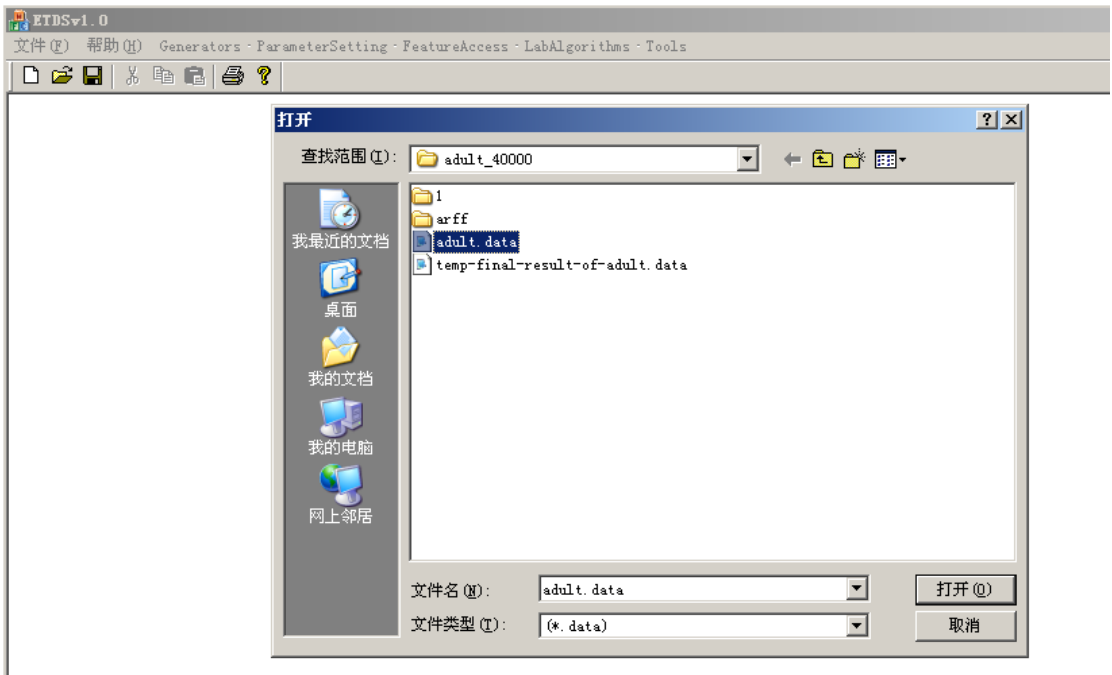
步骤(a) 步骤 2-1—选择 Read.names file 按钮，弹出右侧文件打开框



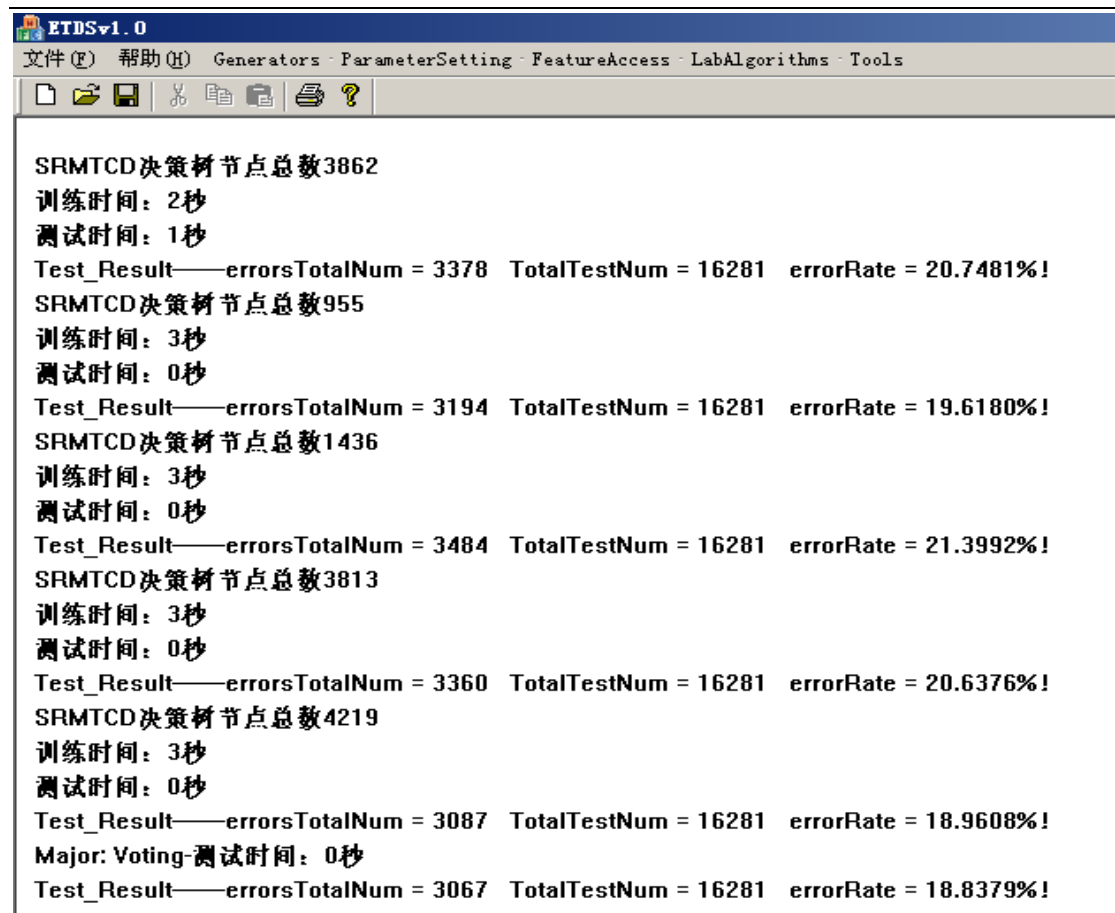
(b) 步骤 2-2—选择步骤 2-1 右侧打开窗口的打开按钮，读取数据库特征并显示



(c) 步骤 3-1—选择 SRMTCD 算法，弹出打开对话框（如下图）



(d) 步骤 3-2—打开训练数据集 adult.data



(e) SRMTCDS 算法运行完毕时的分类结果（以 adult 数据库为例）

图 10.26 SRMTCDS 算法运行中的特征读取、算法运行以及结果显示

10.6 EDT 算法

与 SRMTCDS 与 SRMTCDS 算法运行步骤相似，EDT 算法的运行也分为三大步骤，步骤 1：相关参数的设置；步骤 2：数据库特征文件的读取；步骤 3：算法运行。以下将针对界面操作逐步介绍算法运行的三大步骤。

10.6.1 EDT 算法参数设定菜单

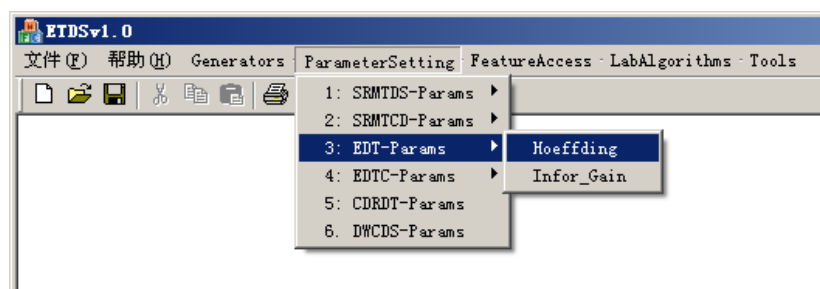


图 10.27 EDT 算法参数设定主菜单

由图 10.27 可知，EDT 算法中参数设定按照分割方法不同分为 Hoeffding 方式与

Infor_Gain 两种方式，选择两种方式的菜单弹出相应的参数对话框，不同在于后者对应的参数设定中没有 Hoeffding bounds 不等式中的一些参数，如： δ 与 τ ，此种情况与算法 SRMTDS 与 SRMTCD 一致。

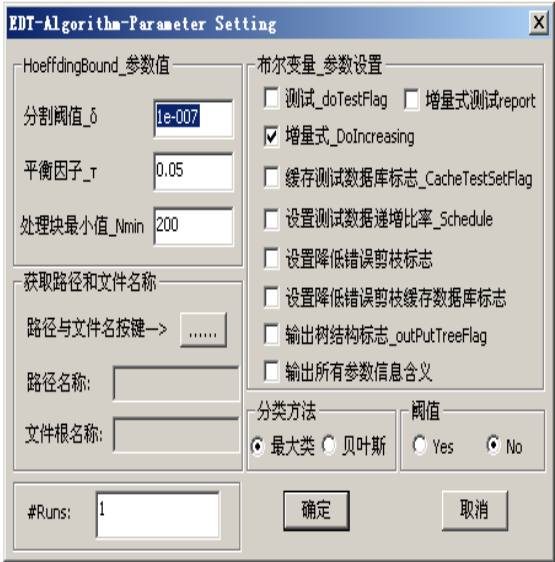


图 10.28 选择 Hoeffding 菜单弹出的对话框

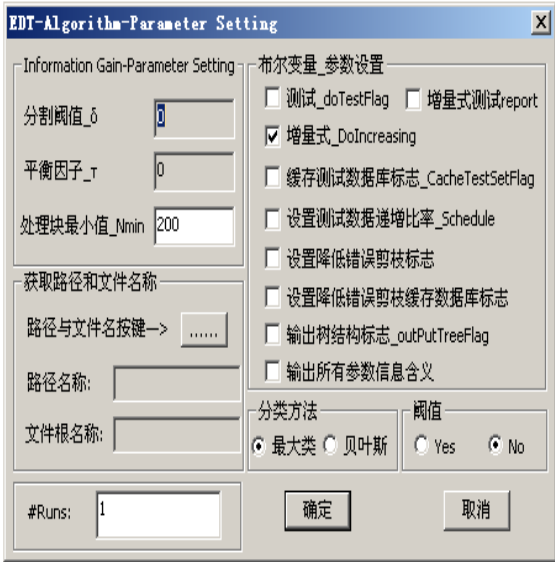


图 10.29 选择 Infor_Gain 菜单弹出的对话框

现将参数列举如下，见表 10.8，由于 EDT 算法的参数设置界面与 SRMTDS 算法的相同，故此表中列出的参数与表 10.4 类似。

表 10.8 EDT 算法一般参数描述（针对图 10.28 与图 10.29）

参数名称	参数意义	类型	界面对应位置	参数值（初始值）
m_fSplitConfidence	Hoeffding Bounds 不等式中的分割阈值	double	分割阈值_δ	10-7
m_fTieConfidence	避免 ties 的阈值	double	平衡因子_τ	0.05
m_iMinChunk	决策树结点中运行一次分割所需的数据量大小	int	处理块最小值_Nmin	200
~	设置生成数据库的路径，读取.test 文件	string	路径与文件名按键->	无默认值
m_ShowSourceDirctoryInEdit	只读项，显示读取的文件名称路径	string	路径名称	
m_ShowFileStemInEdit	只读项，显示读取的.test 文件名称的无后缀名称，用来指示.data 文件名与此文件根名一致	string	文件根名称	
m_iRuns	指示当前算法运行次数	int	#Runs	1
m_bDoIncrementalReport	增量式输出分类结果	bool	选择框增量式测试	false，即不增量

	标志		report	式输出训练中的分类结果
m_bDoIncre	指示当前算法是批量式处理算法还是增量式处理算法	bool	增量式_DoIncreasing	true; true 表示为增量式算法，false 表示为批处理算法
m_bDoTestFlag	/	bool	选择框测试doTestFlag	false
m_bCacheTestSet	/	bool	缓存测试数据库标志_cacheTestSetFlag	false，即不设置缓冲测试区标志
m_bUseSchedule	/	bool	设置测试数据递增比率_Schedule	false
m_bREPrune	/	bool	设置降低错误剪枝标志	false
m_bCachePruneSet	/	bool	设置降低错误剪枝缓存数据库标志	false
m_bOutPutTrees	/	bool	输出树结构标志m_outPutTreeFlag	false
m_bOutPutHelp	/	bool	输出所有参数信息含义	false
m_bBayesMethod	分类方法	bool	最大类	false，即采用最大类方法
			贝叶斯	
m_bConValuesThres	阈值	bool	Yes	false，一般在选择采用贝叶斯分类方法时，此参数才设置为有效状态
			No	

注：m_bCacheTestSet，m_bUseSchedule，m_bREPrune，m_bCachePruneSet，m_bOutPutTrees，m_bOutPutHelp 等变量主要针对经典 VFDTc 算法，在此处没有特别用处。m_bDoIncrementalReport，m_bConValuesThres 值为 true 时，即处于有效状态时，会相应弹出图 10.30 与图 10.31 的对话框，相应设置增量式输出的事例间隔参数 m_increReportCount（默认值为 10000）与连续属性处理时累积的连续属性值个数阈值 m_maxNumOfConValues（默认值为 1000）。

假设 EDT 算法此时采用 Hoeffding Bounds 不等式作为决策树增长结点分割方法见图

10.28 界面，在此界面中，选择  按钮，打开***.test 文件（以 adult.test）为例，故以下的操作均是在以选择 adult.test 文件之后的界面交互操作。

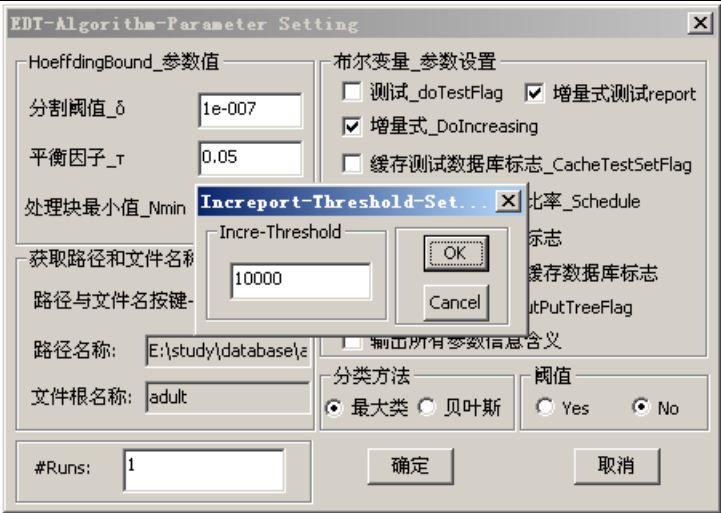


图 10.30 选择增量式测试 report 弹出的对话框

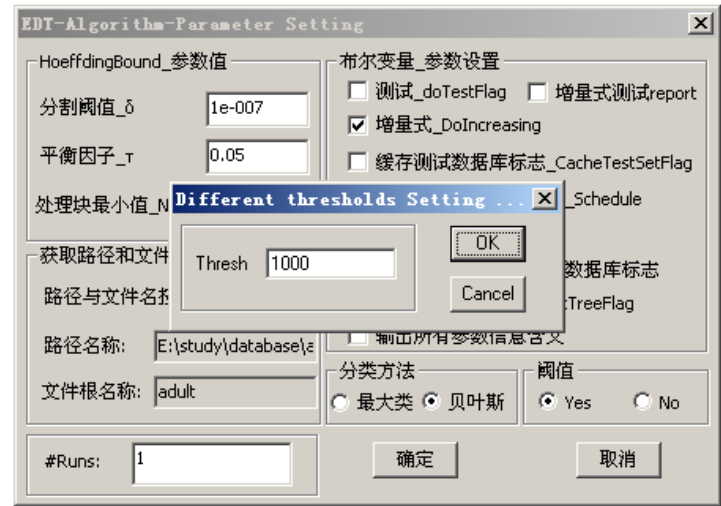


图 10.31 选择贝叶斯分类方法与阈值设定后弹出的对话框

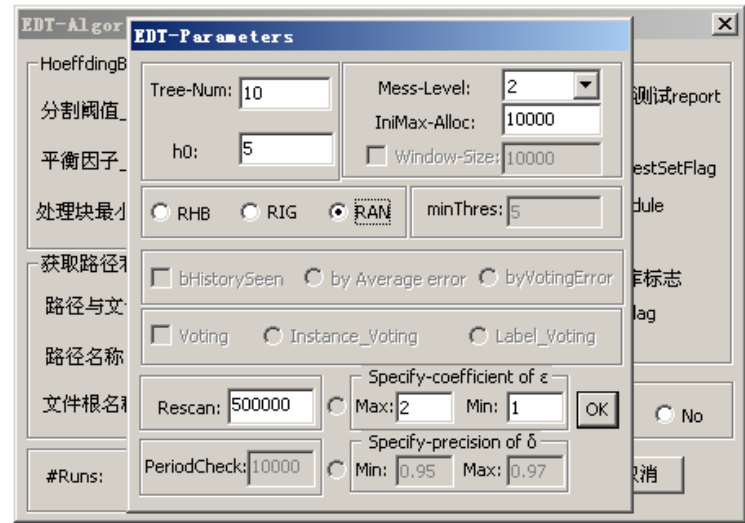


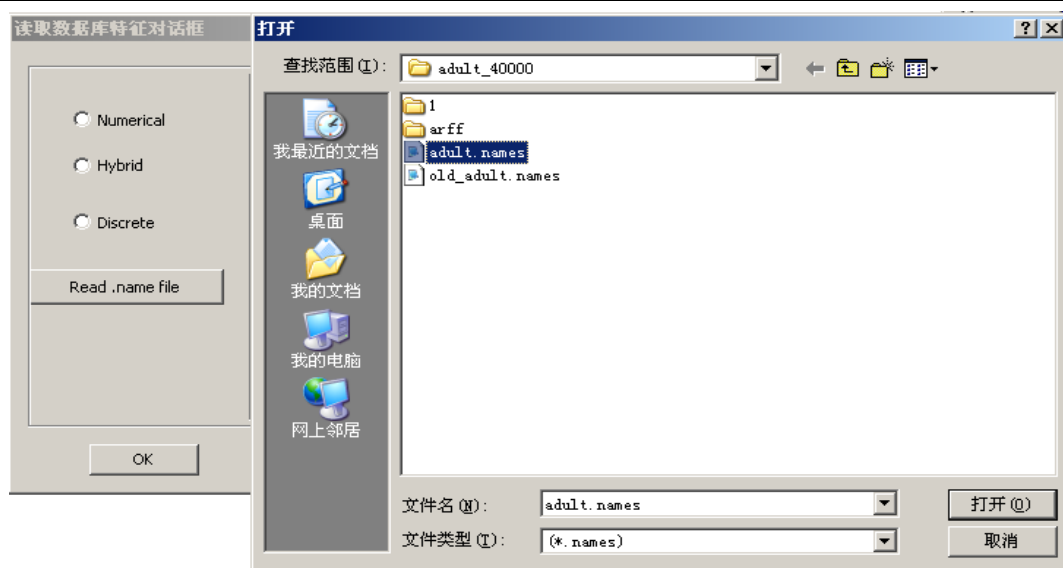
图 10.32 选择图 10.31 中确定按钮后弹出 EDT 特有参数设定对话框(人工选择 RAN 分割测试方式)

表 10.9 EDT 算法参数描述（针对图 10.32）

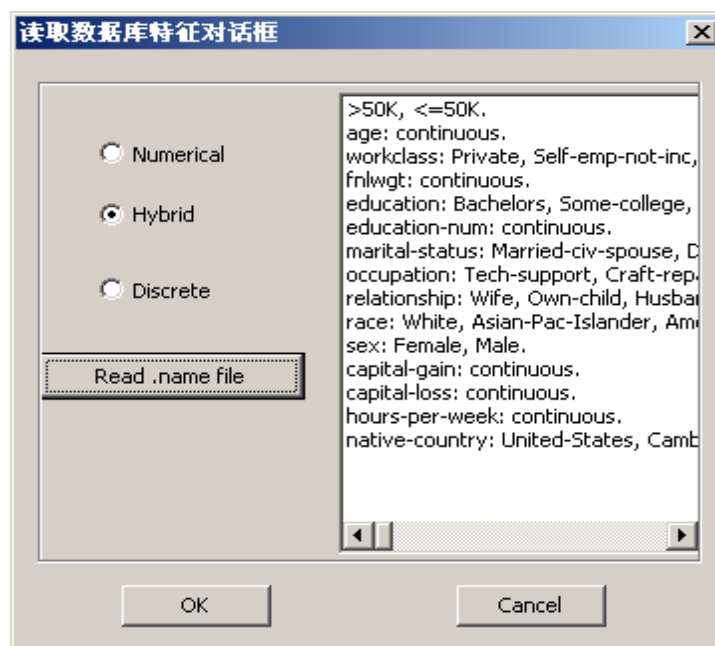
参数名称	参数意义	类型	界面对应位置	参数值（初始值）
m_iTreeNum	集成分类器中决策树构建的棵树	int	Tree-Num	10
m_initHeightOfTree	决策树的树高阈值	int	h_0	5
/	输出信息阈值	int	Mess-Level	2
/	内存检测阈值	int	IniMax-Alloc	
m_iSplitRandomType	分割方法	enum	RHB RIG RAN	HOEFFDINGSELECTION, 即采用半随机策略分割阈值, RHB 单选框选中
m_iRescanPeriod	重新扫描周期	int	Rescan	500000
m_maxCoefficient	概念漂移检测中系数最大值	int	Max	2
m_minCoefficient	概念漂移检测中系数最小值	int	Min	1
	确定此对话框参数设置有效的按钮		OK 按键	

10.6.2 EDT 算法特征数据库读取与算法运行菜单

以上涉及的参数设置是算法运行的第一步骤，完成参数设置之后，需要进行以下两步骤得以运行相应的算法，即步骤二：读取待处理的数据库的特征（读取 adult.names 文件），见图 10.33-(a)，图 10.33-(b)；步骤三：完成参数设置与特征读取后，选择对应的算法菜单如 EDT 算法，进入选择读取.data 文件的界面，见图 10.33-(c)与图 10.33-(d)，以选择 adult.data 为例，读取了训练集后，选择“打开”按钮，即进入运行 EDT 算法阶段。在算法运行阶段，界面无任何提示状态，一旦算法运行完毕，相应的分类结果在界面显示，如图 10.33-(e)。前面 10 个 Test_Result 对应的行表示 EDT 算法中构建的集成分类器各自的分类效果，最后一个 Test_Result 对应的行是 EDT 算法中构建的集成分类器的最终投票结果，即算法的分类结果。



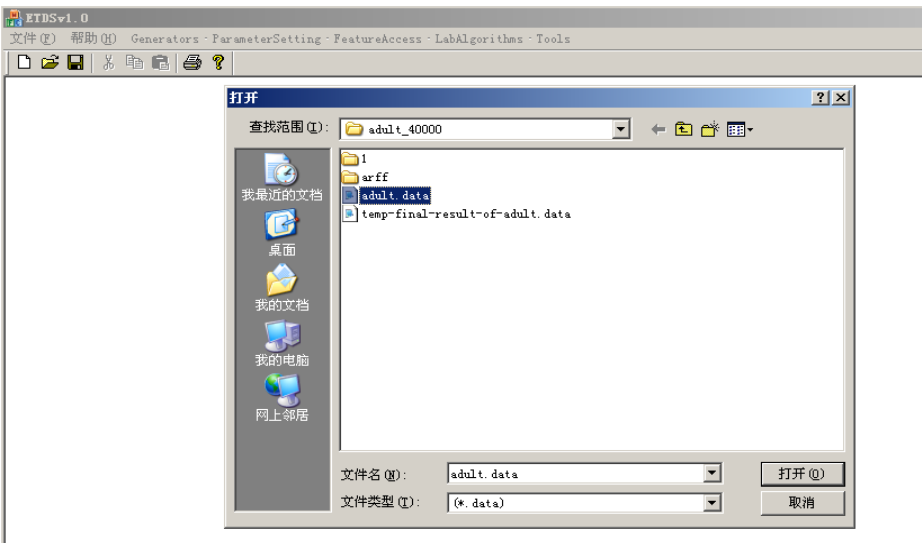
(a) 步骤 2-1—选择 Read.names file 按钮，弹出右侧文件打开框



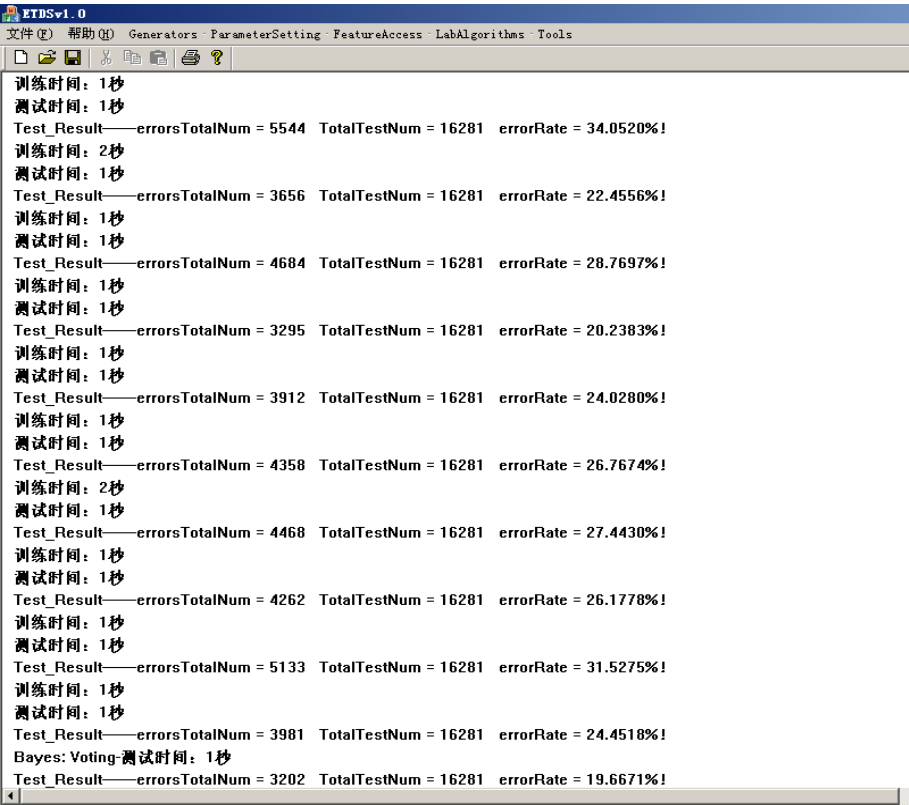
(b) 步骤 2-2—选择步骤 2-1 右侧打开窗口的打开按钮，读取数据库特征并显示



(c) 步骤 3-1—选择 EDT 算法，弹出打开对话框（如下图）



(d) 步骤 3-2—打开训练数据集 adult.data



(e) EDT 算法运行完毕时的分类结果(以 adult 数据库为例)

图 10.33 EDT 算法运行中的特征读取、算法运行以及结果显示

10.7 EDTC 算法

作为 EDT 算法的扩展算法，EDTC 算法的运行也分为三大步骤，步骤 1：相关参数的设置；步骤 2：数据库特征文件的读取；步骤 3：算法运行。以下将针对界面操作逐步介绍

算法运行的三大步骤。

10.7.1 EDTC 算法参数设定菜单

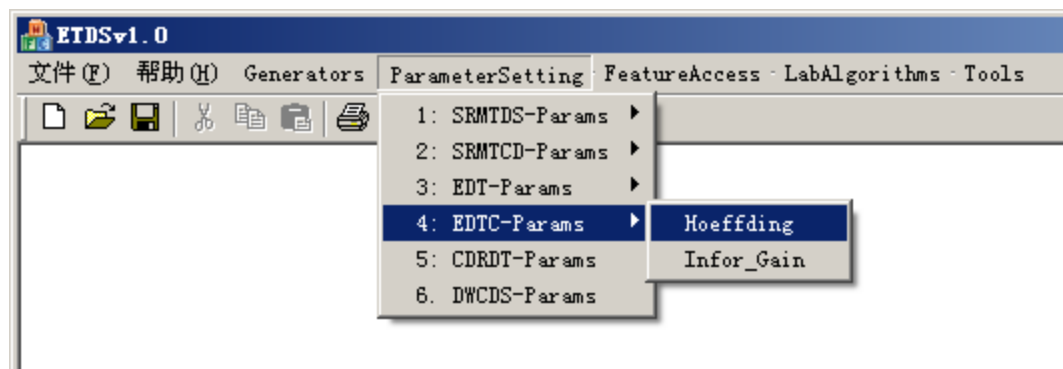


图 10.34 EDTC 算法参数设定主菜单

由图 10.34 可知, EDTC 算法中参数设定按照分割方法不同分为 Hoeffding 方式与 Infor_Gain 两种方式, 选择两种方式的菜单弹出相应的参数对话框(如图 10.35 和图 10.36 所示), 不同在于后者对应的参数设定中没有 Hoeffding bounds 不等式中的一些参数, 如: δ 与 τ , 此种情况与算法 SRMTDS、SRMTCD 与 EDT 一致。

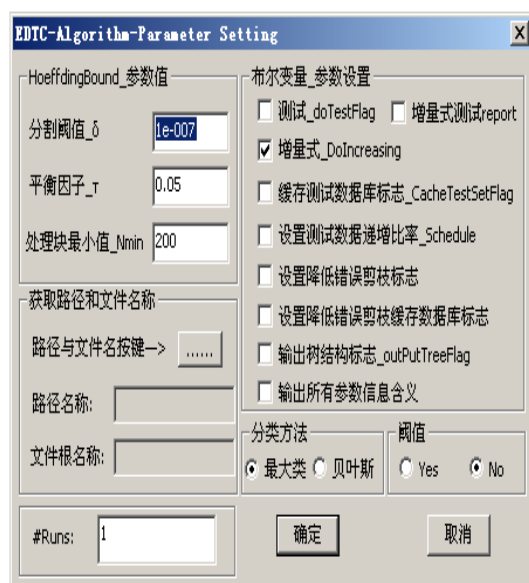


图 10.35 选择 Hoeffding 菜单弹出的对话框



图 10.36 选择 Infor_Gain 菜单弹出的对话框

图 10.35 对应的参数列表与表 10.8 同, 故此处不再重复列出。需要指出的是选中界面中复选框“增量式测试 report”, 弹出阈值设定对话框, 如图 10.37。而选中“贝叶斯”分类方法单选框与阈值设定框中的“**Yes**”单选框, 则弹出图 10.38 所示的阈值设定对话框。注: 图 10.37 与图 10.38 均以读取 adult 数据库为例。

假设 EDTC 算法此时采用 Hoeffding Bounds 不等式作为决策树增长结点分割方法见图

10.35 界面,在此界面中,选择 **路径与文件名按钮->** 按钮,打开***.test 文件(以 adult.test)为例,故以下的操作均是在以选择 adult.test 文件之后的界面交互操作。

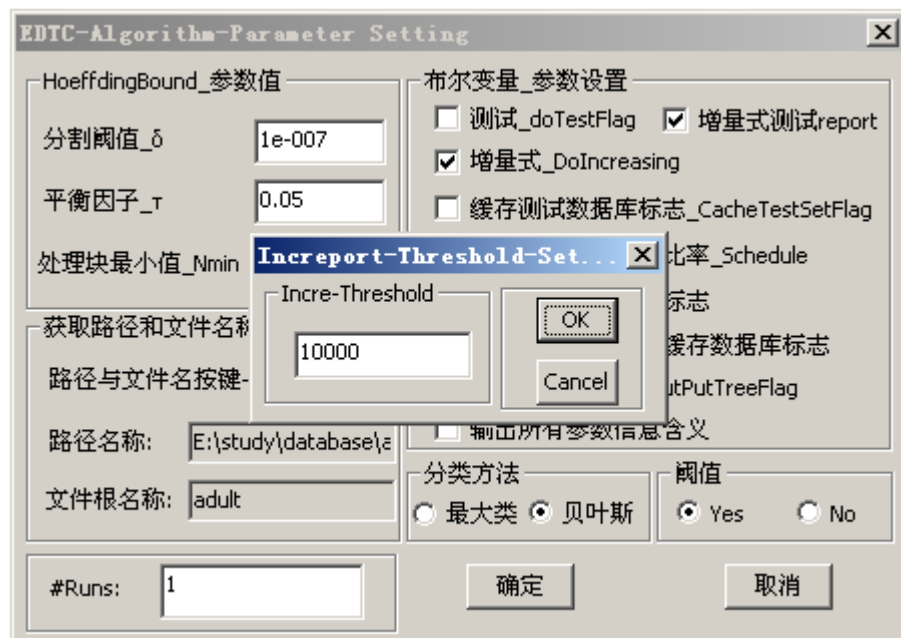


图 10.37 选择增量式测试 report 弹出的对话框

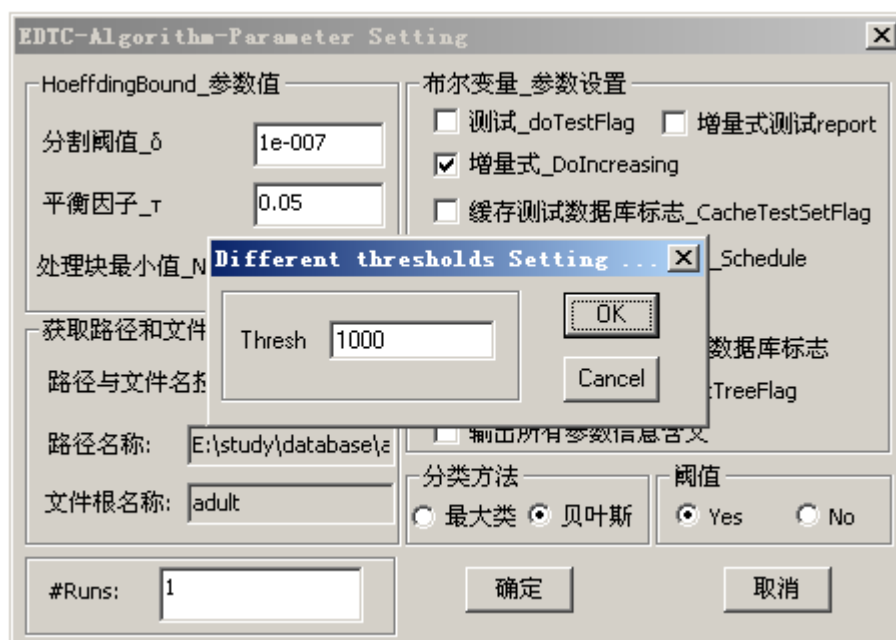


图 10.38 选择贝叶斯分类方法与阈值设定后弹出的对话框

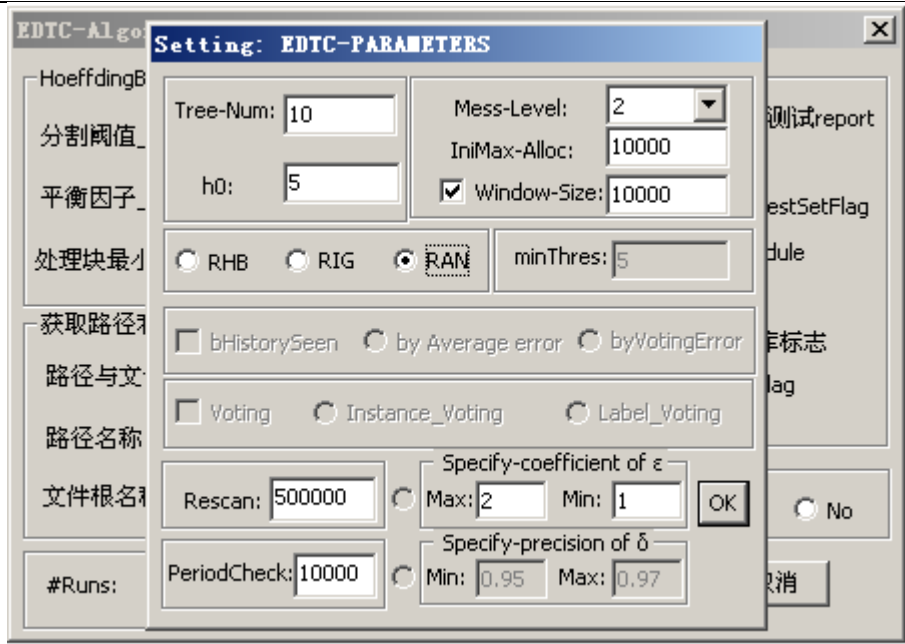


图 10.39 选择图 10.38 中确定按钮后弹出 EDTC 特有参数设定对话框(人工选择 RAN 分割测试方式)

针对图 10.39 对话框参数描述如下，见表 10.10。

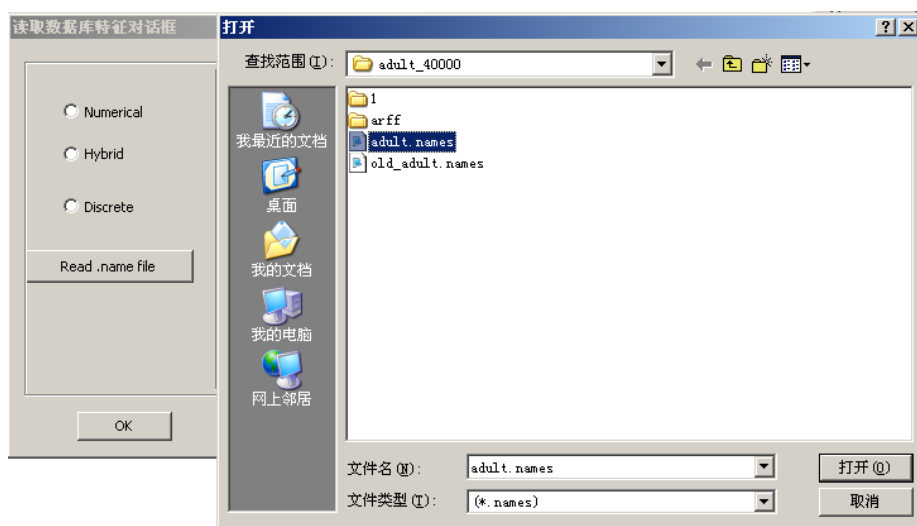
表 10.10 EDTC 算法一般参数描述（针对图 10.39）

参数名称	参数意义	类型	界面对应位置	参数值（初始值）
m_iTreeNum	集成分类器中决策树构建的棵树	int	Tree-Num	5
m_initHeightOfTree	决策树的树高阈值	int	h_0	5
/	输出信息阈值	int	Mess-Level	2
/	内存检测阈值	int	IniMax-Alloc	
m_iSplitRandomType	分割方法	enum	RHB RIG RAN	HOEFFDINGSELECTION，即采用半随机策略分割阈值，RHB 单选框选中
m_iRescanPeriod	重新扫描周期	int	Rescan	500000
m_iCheckPeriod	概念漂移检测周期	int	PeroidCheck	10000
m_maxCoefficient	概念漂移检测中系数最大值	int	Max	2
m_minCoefficient	概念漂移检测中系数最小值	int	Min	1

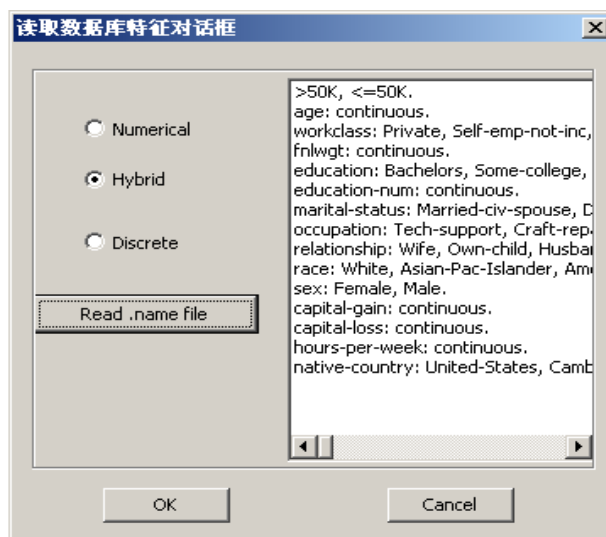
10.7.2 EDTC 算法特征数据库读取与算法运行菜单

以上涉及的参数设置是算法运行的第一步骤，完成参数设置之后，需要进行以下两步骤得以运行相应的算法，即步骤二：读取待处理的数据库的特征（读取 adult.names 文件），见

图 10.40-(a)，图 10.40-(b)；步骤三：完成参数设置与特征读取后，选择对应的算法菜单如 EDT 算法，进入选择读取.data 文件的界面，见图 10.40-(c)与图 10.40-(d)，以选择 adult.data 为例，读取了训练集后，选择“打开”按钮，即进入运行 EDTC 算法阶段。在算法运行阶段，界面无任何提示状态，一旦算法运行完毕，相应的分类结果在界面显示，如图 10.40-(e)。前面 10 个 Test_Result 对应的行表示 EDTC 算法中构建的集成分类器各自的分类效果，最后一个 Test_Result 对应的行是 EDTC 算法中构建的集成分类器的最终投票结果，即分类效果。



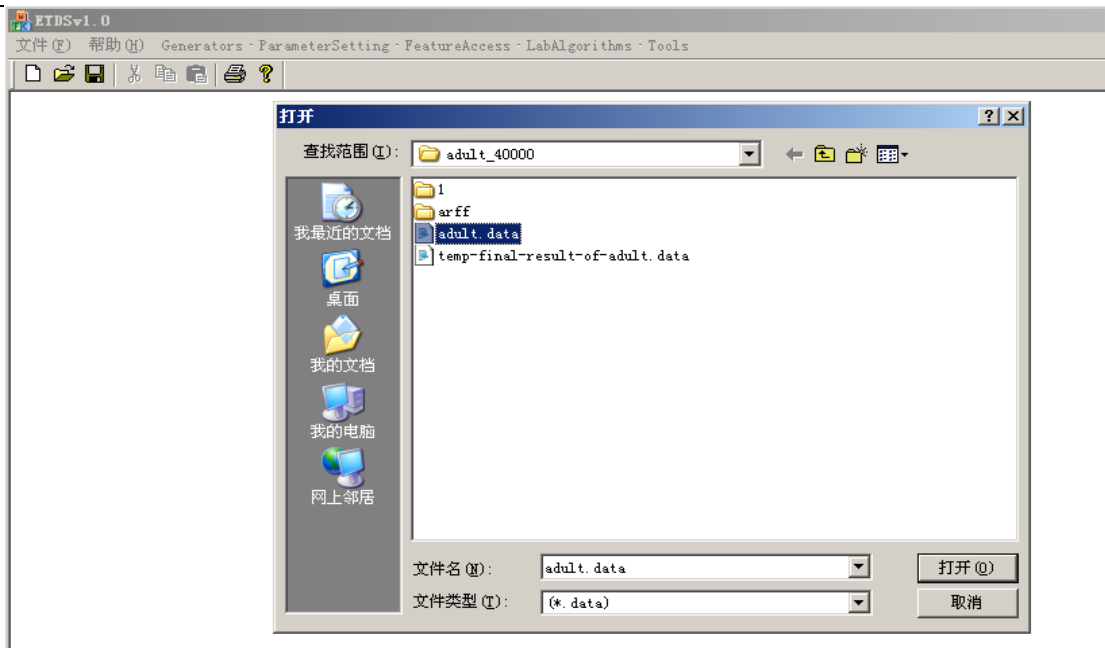
(a) 步骤 2-1—选择 Read.names file 按钮，弹出右侧文件打开框



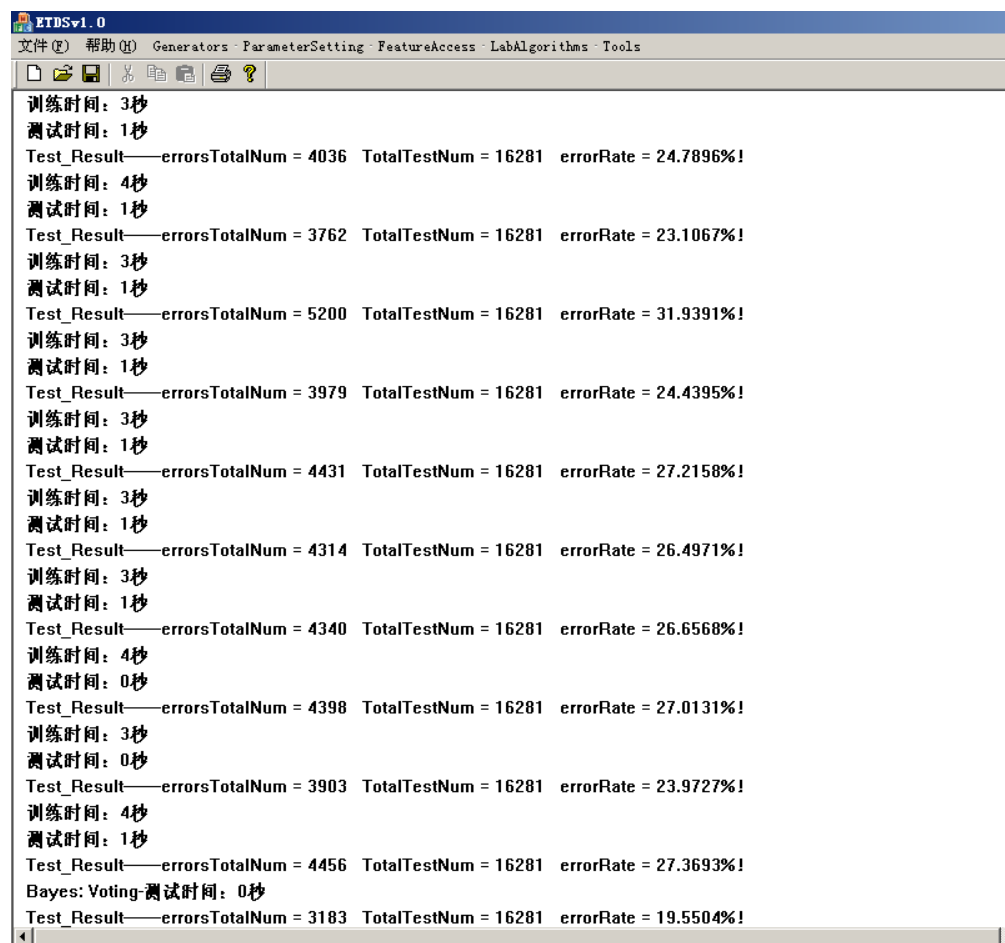
(b) 步骤 2-2—选择步骤 2-1 右侧打开窗口的打开按钮，读取数据库特征并显示



(c) 步骤 3-1—选择 EDTC 算法，弹出打开对话框（如下图）



(d) 步骤 3-2—打开训练数据集 adult.data



(e) (e) EDTC 算法运行完毕时的分类结果(以 adult 数据库为例)

图 10.40 EDTC 算法运行中的特征读取、算法运行以及结果显示

10.8 CDRDT 算法

与 EDTC 算法相似，CDRDT 算法的运行也分为三大步骤，步骤 1：相关参数的设置；步骤 2：数据库特征文件的读取；步骤 3：算法运行。以下将针对界面操作逐步介绍算法运行的三大步骤。

10.8.1 CDRDT 算法参数设定菜单

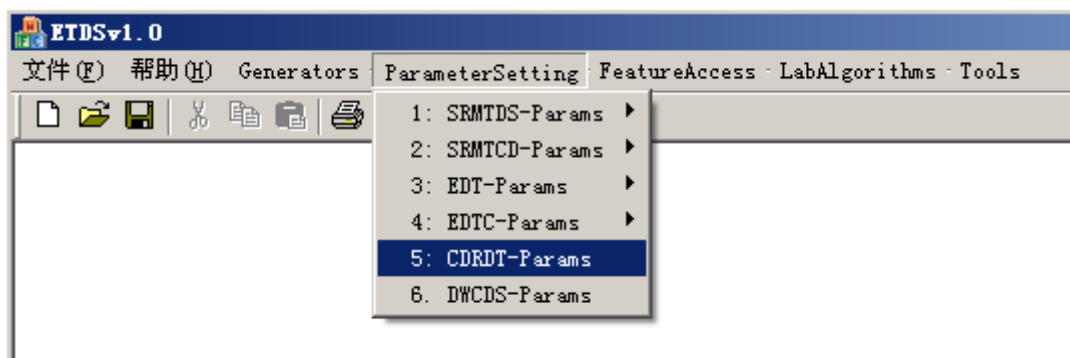


图 10.41 CDRDT 算法参数设定主菜单

如图 10.41，选择 CDRDT-Params 菜单即可进入 CDRDT 算法中一般参数设置界面，见图 10.42。由此图可知，CDRDT 一般参数设置界面与 EDT/EDTC 选择 Hoeffding 分割方式进入的界面相似。



图 10.42 CDRDT 算法中一般参数设定对话框

图 10.42 对应的参数列表与表 10.8 同，故此处不再重复列出。在图 10.42 中，选择 **路径与文件名按键->** 按钮，打开***.test 文件（以 adult.test）为例，故以下的操作均是在以选择 adult.test 文件之后的界面交互操作。另外，需要指出的是选中界面中复选框“增

量式测试 report”，弹出阈值设定对话框，如图 10.43。而选中“贝叶斯”分类方法单选框与阈值设定框中的“**Yes**”单选框，则弹出图 10.44 所示的阈值设定对话框。图 10.43 与图 10.44 和以上各章节相应部分的图形相似，不同在于界面名称的变化。

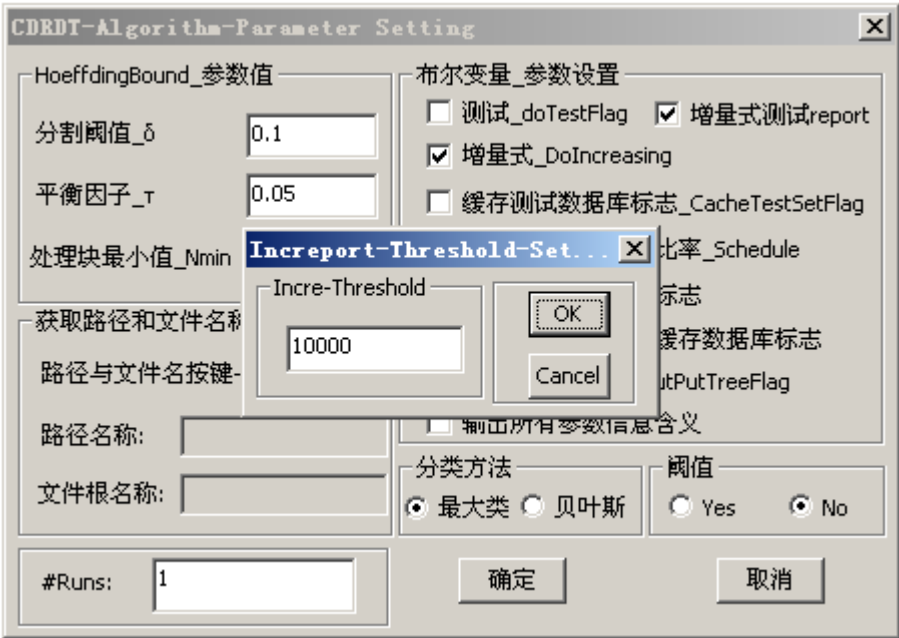


图 10.43 选择增量式测试 report 弹出的对话框

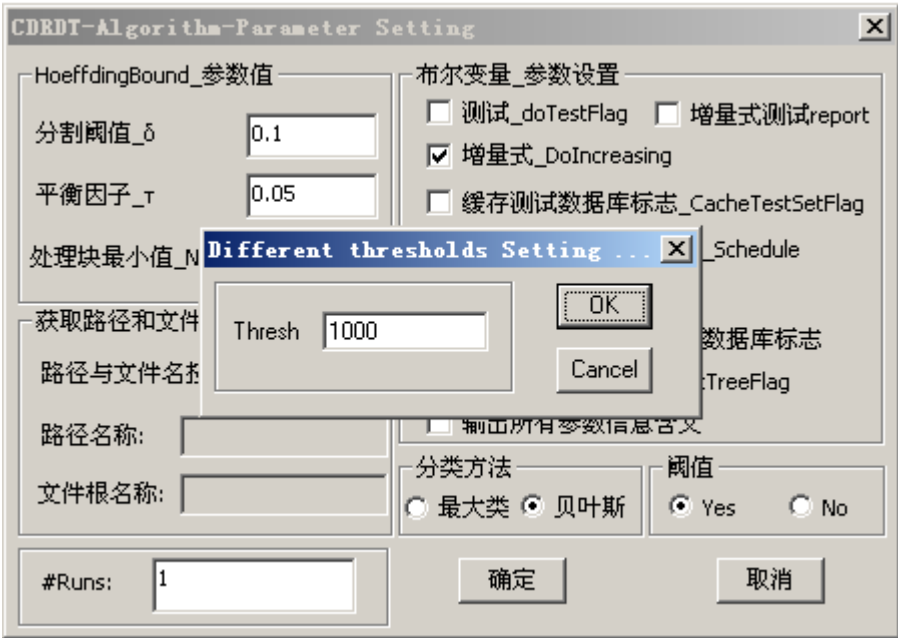


图 10.44 选择贝叶斯分类方法与阈值设定后弹出的对话框

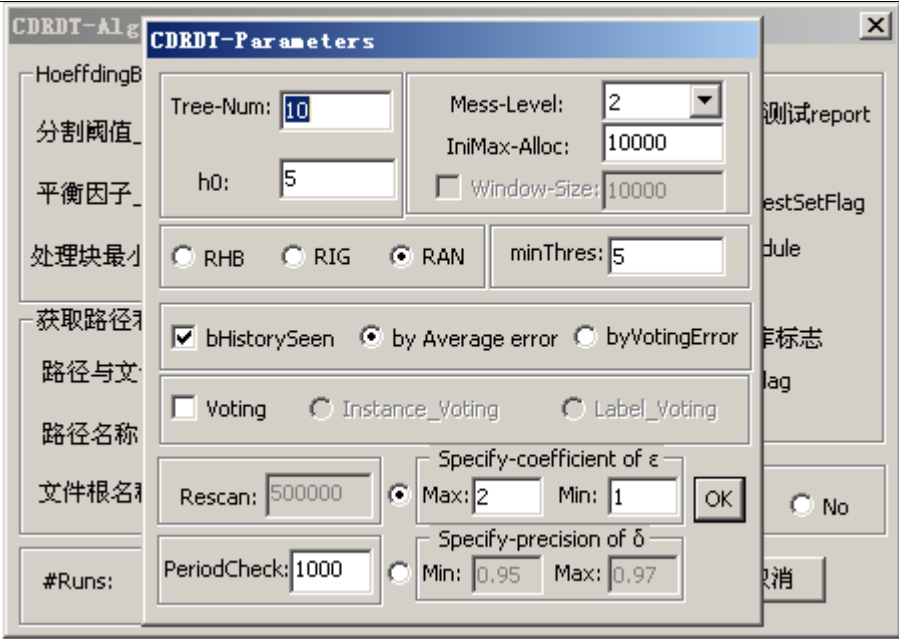


图 10.45 选择图 10.44 中确定按钮后弹出 CDRDT 特有参数设定对话框

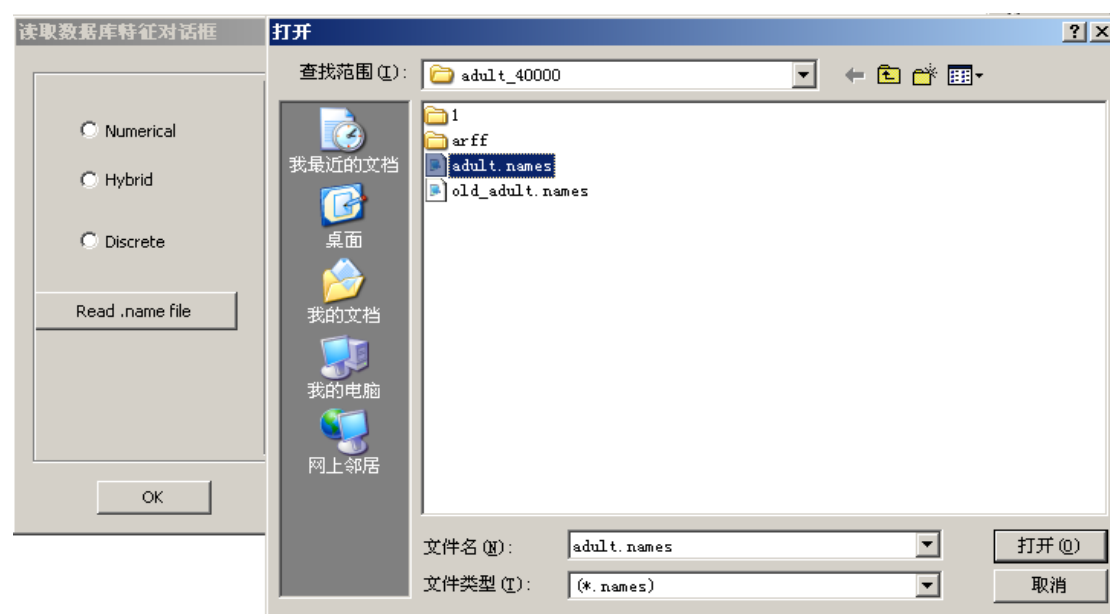
针对图 10.45 对话框参数描述如下，见表 10.11。

表 10.11 CDRDT 算法一般参数描述（针对图 10.45）

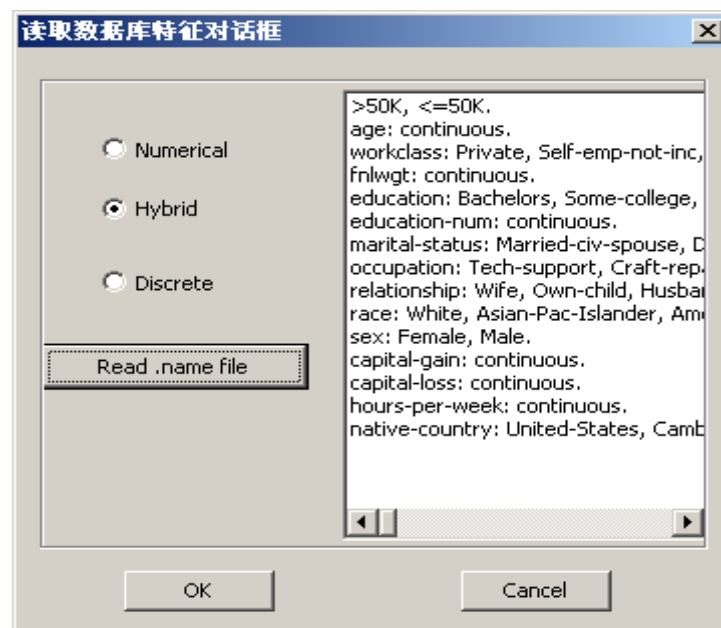
参数名称	参数意义	类型	界面对应位置	参数值（初始值）
m_iTreeNum	集成分类器中决策树构建的棵树	int	Tree-Num	5
m_initHeightOfTree	决策树的树高阈值	int	h0	5
/	输出信息阈值	int	Mess-Level	2
/	内存检测阈值	int	IniMax-Alloc	
m_iSplitRandomType	分割方法	enum	RHB RIG RAN	RANDOMSELECTION，即采用完全随机策略分割阈值，RAN 单选框选中
m_minThres	用户 CDRDT 类中 ComputeErrorRateByBlockSeen 函数中计算错误率时的阈值，即事例大于此阈值才计算错误率	int	minThres	5
m_bHistorySeen	计算错误率时是否利用历史数据	布尔值	bHistorySeen	true*
m_iCheckPeriod	概念漂移检测周期	int	PeroidCheck	1000
m_maxCoefficient	概念漂移检测中系数最大值 ϵ 的系数	int	Max	2
m_minCoefficient	概念漂移检测中系数最小值 ϵ 的系数	int	Min	1
m_maxTheta	概念漂移检测中系数最大值 δ 的系数	double	Max	0.97

m_minTheta	概念漂移检测中系数最小值 δ 的系数	double	Min	0.95
	此复选框在 CDRDT 中是欲留参数		Voting	
<p>* 勾选复选框 bHistorySeen: 实验证明利用历史数据有助于提高模型的分类效果;</p> <p>另外, 此界面中 ε 与 δ 的系数设置是互斥的, 即在实验过程中, 要么 Specify-coefficient of ε 有效要么 Specify-precision of δ 有效。</p>				

10.8.2 CDRDT 算法特征数据库读取与算法运行菜单



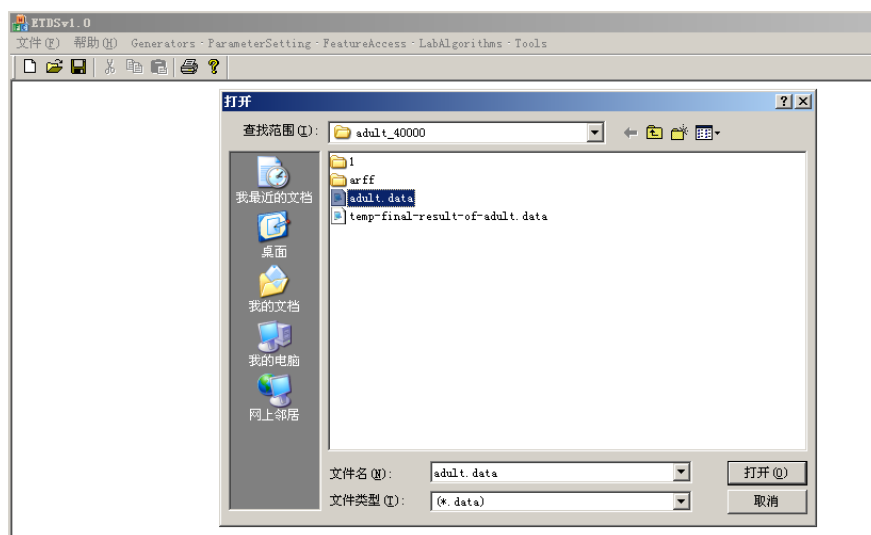
(a) 步骤 2-1—选择 Read.names file 按钮, 弹出右侧文件打开框



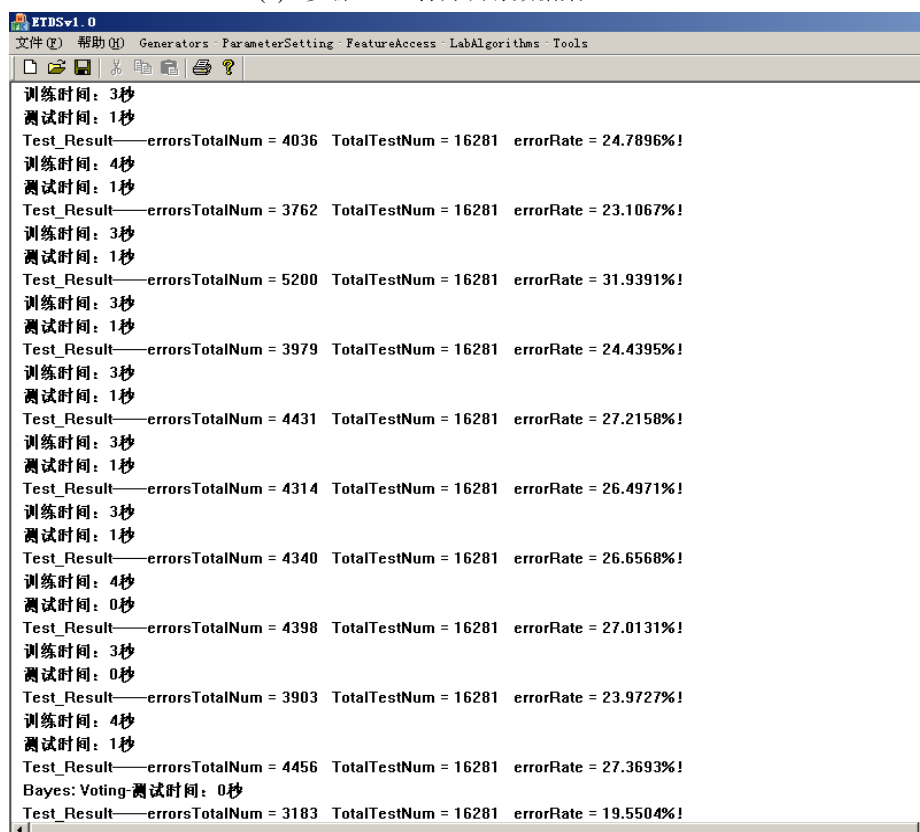
(b) 步骤 2-2—选择步骤 2-1 右侧打开窗口的打开按钮, 读取数据库特征并显示



(c) 步骤 3-1—选择 CDRDT 算法，弹出打开对话框（如下图）



(d) 步骤 3-2—打开训练数据集 adult.data



(e) CDRDT 算法运行完毕时的分类结果（以 adult 数据库为例）

图 10.46 CDRDT 算法运行中的特征读取、算法运行以及结果显示

以上涉及的参数设置是算法运行的第一步骤，完成参数设置之后，需要进行以下两步骤得以运行相应的算法，即步骤二：读取待处理的数据库的特征（读取 adult.names 文件），见图 10.46-(a)，图 10.46-(b)；步骤三：完成参数设置与特征读取后，选择对应的算法菜单如 EDT 算法，进入选择读取.data 文件的界面，见图 10.46-(c)与图 10.46-(d)，以选择 adult.data 为例，读取了训练集后，选择“打开”按钮，即进入运行 CDRDT 算法阶段。在算法运行阶段，界面无任何提示状态，一旦算法运行完毕，相应的分类结果在界面显示，如图 10.46-(e)。前面 10 个 Test_Result 对应的行表示 CDRDT 算法中构建的集成分类器各自的分类效果，最后一个 Test_Result 对应的行是 CDRDT 算法中构建的集成分类器的最终投票结果，即算法的分类效果。

10.9 DWCDS 算法

与 CDRDT 算法相似，DWCDS 算法的运行也分为三大步骤，步骤 1：相关参数的设置；步骤 2：数据库特征文件的读取；步骤 3：算法运行。以下将针对界面操作逐步介绍算法运行的三大步骤。

10.9.1 DWCDS 算法参数设定菜单

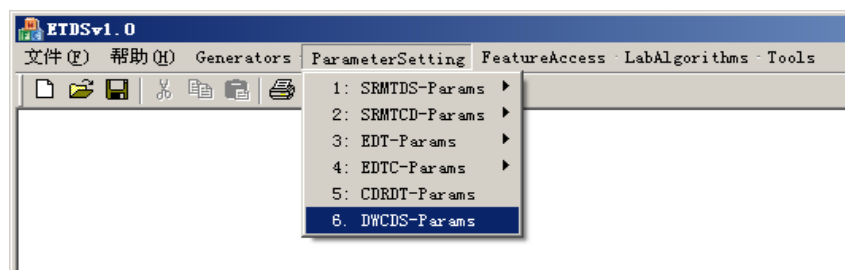


图 10.47 DWCDS 算法参数设定主菜单

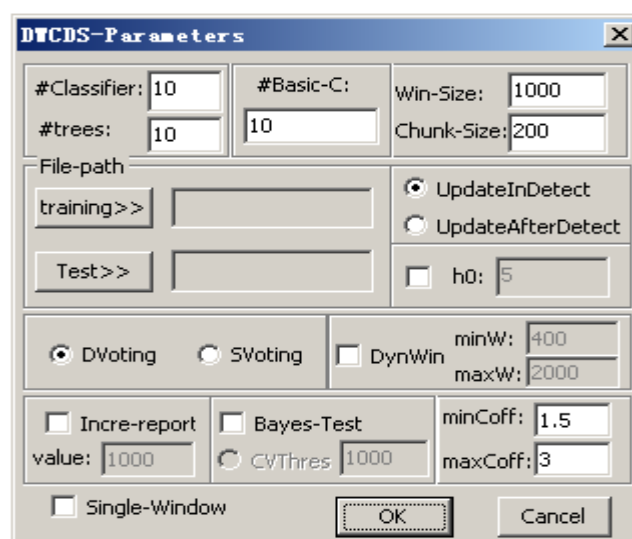


图 10.48 DWCDS 算法中一般参数设定对话框

如图 10.47, 选择 DWCDs-Params 菜单即可进入 DWCDs 算法中一般参数设置界面, 见图 10.48。

图 10.48 对应的参数列表见表 10.12。

表 10.12 DWCDs 算法参数描述 (针对图 10.48)

参数名称	参数意义	类型	界面对应位置	参数值 (初始值)
m_iClassifierCount	集成分类器中基分类器的个数	int	# Classifier	10
m_treeCount	每个基分类器中决策树的个数, 最大棵树为 10 棵 1)	int	# trees	10
m_iBasicClassifier	初始建立基分类器的个数	int	# Basic-C	10
m_iWinSize	滑动窗口的初始大小	int	Win-Size	1000
m_iChunkSize	基本窗口的大小	int	Chunk-Size	200
m_bUpdateInDetec	值为 true 时: 算法在检测到漂移点之前对分类器进行更新, 即 UpdateInDetec 单选框被勾选; 则表示算法在检测到漂移点以后才对分类器进行更新, 即 UpdateAfterDetect 单选框被勾选	bool	UpdateInDetec 与 UpdateAfterDetect	true
	树的最大高度 2)	int	h0	5
m_trainFileShow	读入并显示训练数据所在路径	string	training>>	NULL
m_testFileShow	读入并显示测试数据所在路径	string	Test	NULL
m_bDVoting	表示双层投票机制的标志 3)	bool	DVoting 与 SVoting	true
m_bDynamicWin	DWCDs 算法中采用动态调整窗口机制的标志 4)	bool	DynWin	false
m_minWindow	滑动窗口的最小值	int	minW	
m_maxWindow	滑动窗口的最大值	int	maxW	
m_bIncreReport	对测试结果采用增量式的输出过程	bool	Incre-report	false
m_increReportCount	增量式输出结果的事例间隔阈值	int	value	false
m_bayesTest	在决策树的叶子节点采用 NB 方法进行分类; 否则为最大类方法	bool	Bayes-Test	false
m_maxContValThres	采用 Bayes-Test 方式时每个叶子节点所存储的最大事例个数 5)	int	CVThres	1000
m_minCoeff	采用伯努利分布方法进行漂移检测时的警告系数	double	minCoff	1.5
m_maxCoeff	采用伯努利分布方法进行漂移检测时的漂移系数 6)	double	maxCoff	3
m_bSingleWin	采用单层窗口机制进行漂移检测 7)	bool	Single-Window	false

注：1) 决策树的个数依据公式 10.1 计算， n 表示数据集中属性的维数；

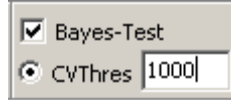
$$K = \min\{C_n^{\lceil n/2 \rceil}, 10\} \quad (\text{式 } 10.1)$$

2) 树的高度的设定一般采用默认值 5，如果用户需要自定义高度，必须选择界面中 h_0 对应的复选框，此时 h_0 的设定才有效；

3) 双层投票机制，即先对每个基分类器中的决策树进行投票，再对 N 个基分类器的结果进行投票；

4) 双层窗口 Win-Size 随漂移检测情况调整窗口大小。具体调整策略为：其中，若检测到概念漂移，滑动窗口 Win-Size 减少一个基本窗口 Chunk-Size 的大小，如果此时滑动窗口 Win-Size 的大小已达到最小阈值 MinSW，则保持原值不变；若没有检测到概念漂移，滑动窗口 Win-Size 则增加一个基本窗口 Chunk-Size 的大小，直至达到最大阈值 MaxSW 时不再增加。在选择了 DynWin 复选框后，m_maxWindow 与 m_minWindow 对应的值可采用用户指定方式，所指定的值就是 DWCDs 算法中窗口调整的最大最小限制；

5) 分割过程中连续属性值个数阈值的设定仅在选择了复选框“Bayes-Test”之后，CVThres 单选框才有效，用户可以选择此单选框从而指定右侧编辑框中的数值，见如下截图：



6) 采用伯努利分布方法进行漂移检测时的警告系数 τ_1 与漂移系数 τ_2 （见公式 10.2、10.3）：

$$P_i + \tau_1 * S_i \leq P_n \quad (\text{式 } 10.2)$$

$$P + \tau_2 * S \leq P' \quad (\text{式 } 10.3)$$

7) 为了对比 DWCDs 算法中设置双窗口机制与单窗口的实验，此界面添加了选择单窗口机制的复选框。实验证明：双窗口机制比单窗口机制更能适应概念漂移。

以 adult 数据库为例，基于图 10.48 界面操作如图 10.49 所示，

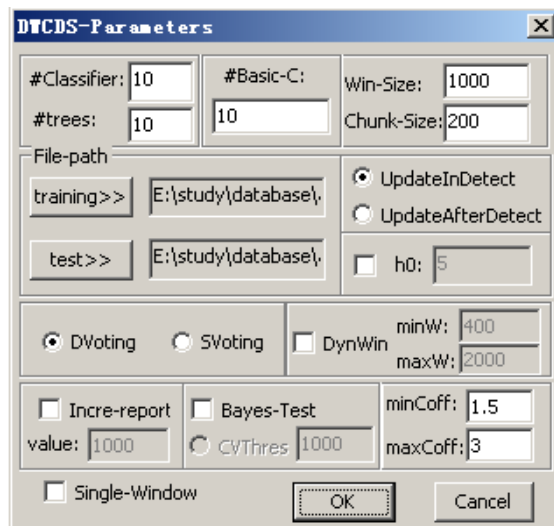
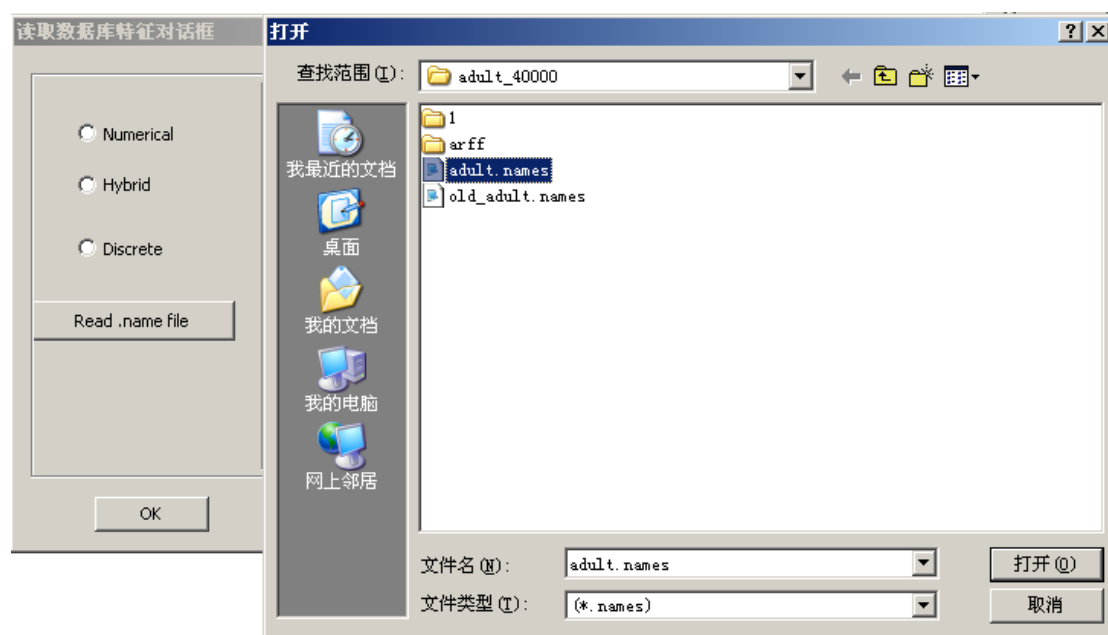


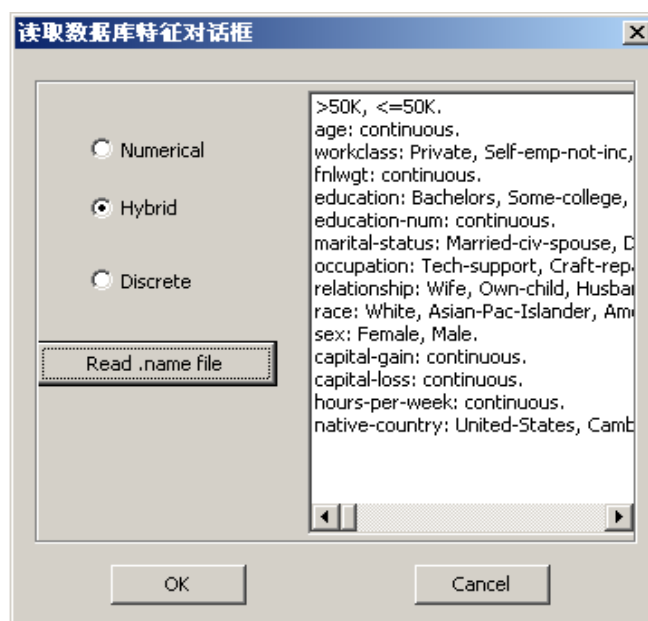
图 10.49 DWCDs 算法运行前参数设置示例

10.9.2 DWCDs 算法特征数据库读取与算法运行菜单

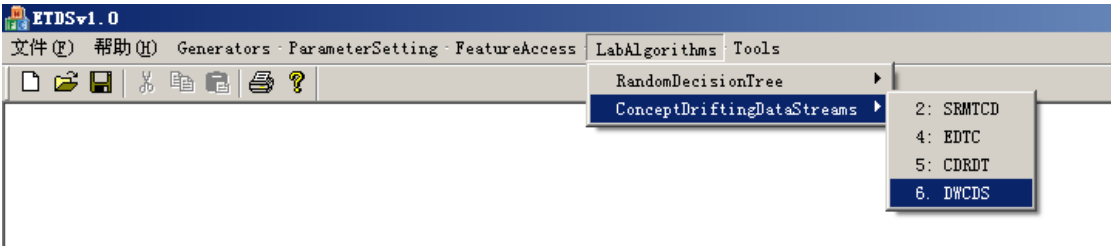
以上涉及的参数设置是算法运行的第一步骤（点击图 10.49 界面中的 OK 按钮即第一步骤设置完成），完成参数设置之后，需要进行以下两步骤得以运行相应的算法，即步骤二：读取待处理的数据库的特征（读取 adult.names 文件），见图 10.50-(a)，图 10.50-(b)；步骤三：完成参数设置与特征读取后，选择对应的算法 DWCDs 菜单，即进入运行 DWCDs 算法阶段。在算法运行阶段，界面无任何提示状态，一旦算法运行完毕，相应的分类结果在界面显示，如图 10.50-(d)，屏幕呈现 DWCDs 算法在测试集上的分类结果。



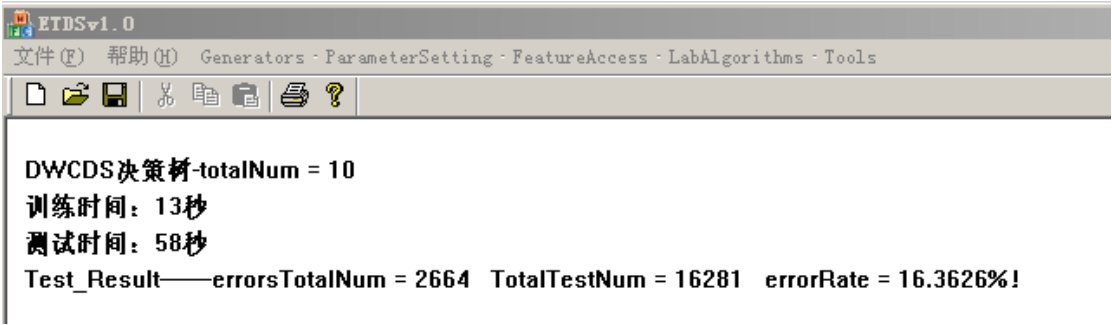
(a) 步骤 2-1—选择 Read.names file 按钮，弹出右侧文件打开框



(b) 步骤 2-2—选择步骤 2-1 右侧打开窗口的打开按钮，读取数据库特征并显示



(c) 步骤 3—选择 DWCDs 算法进入程序运行阶段



(d) DWCDs 算法运行完毕时的分类结果（以 adult 数据库为例）

图 10.50 DWCDs 算法运行中的特征读取、算法运行以及结果显示

注：以上八算法使用示例大部分是在默认参数设置下的实验结果，分类结果除屏幕显示外，更详细的信息会以文件的形式输出，输出路径与用户选择的数据库路径一致。

10.10 布局图与流程图

10.10.1 数据流实验工具算法布局图

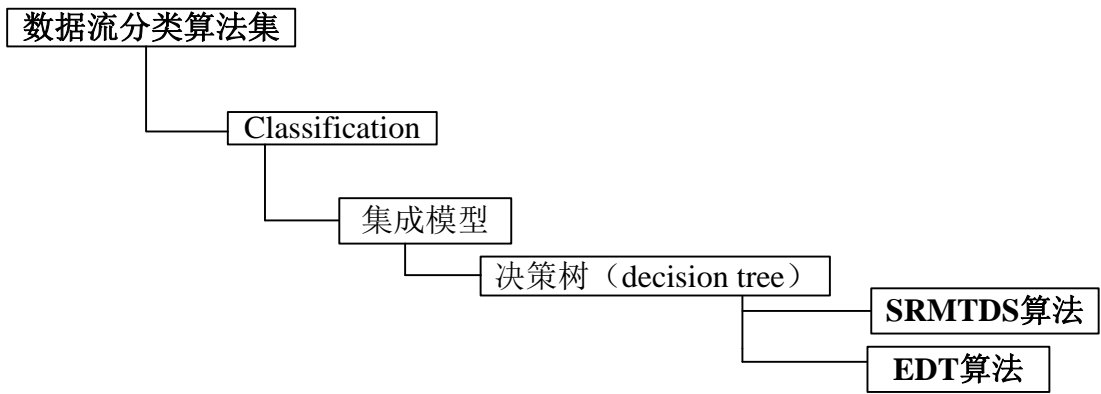


图 10.51 数据流算法布局图

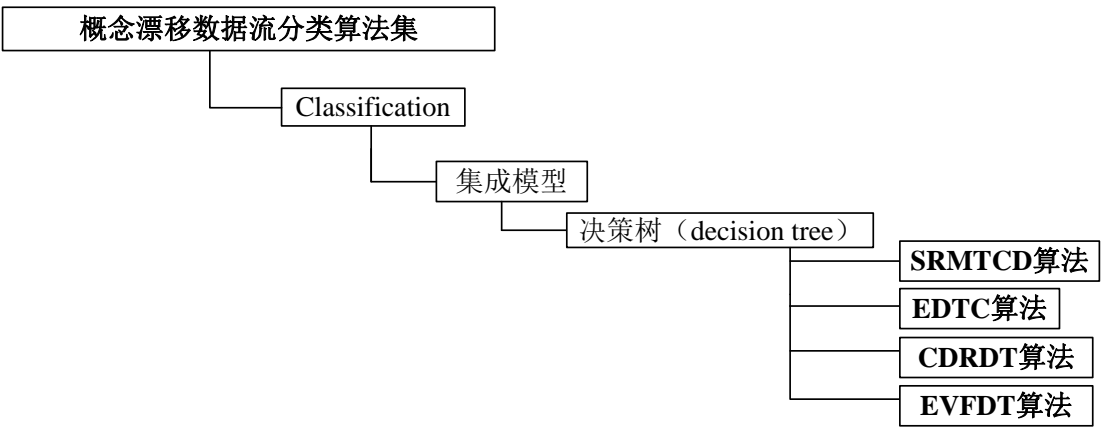


图 10.52 概念漂移数据流算法布局图

10.10.2 数据流分类算法流程图

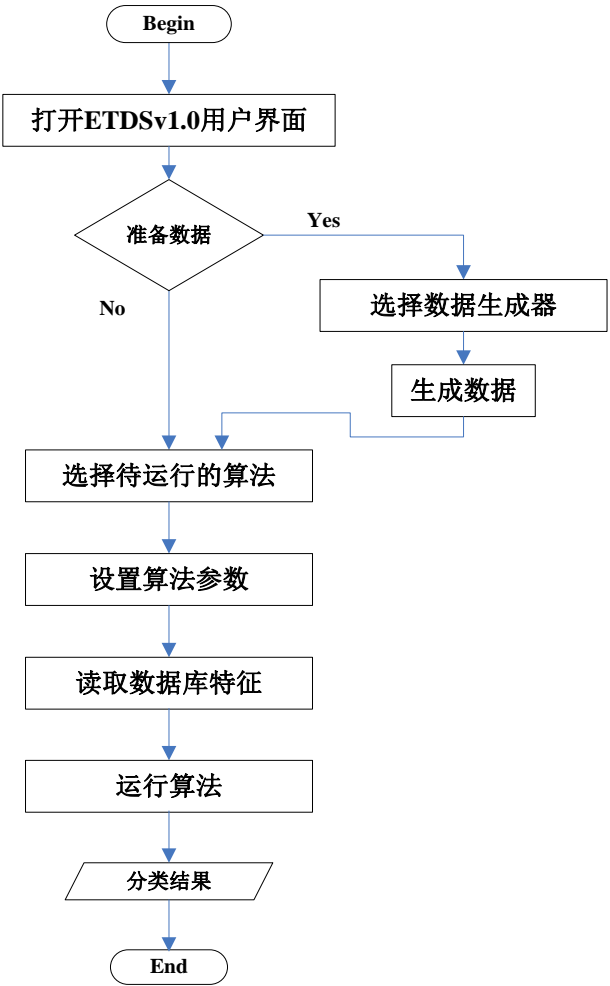


图 10.53 算法流程图

第 11 章 经典的数据流分类算法实验工具

本章主要介绍若干经典的数据流分类算法实验工具的主要功能与使用方法。

11.1 VFML 系统

VFML (Very Fast Machine Learning toolkit) 系统用于挖掘高速数据流与海量数据集。VFML系统包含三大组件：1) 集成了大量的工具与API函数，为用户开发新学习算法提供指南；2) 集成了重要学习算法的实现；3) 集成了可伸缩学习算法（主要由Pedro Domingos 与 GeoffHulten两人开发完成）。VFML系统主要用C语言（少量代码用了Python语言）编码实现，提供了一系列的指南与示例以及丰富的用JAVADoc格式形成的源文档。VFML系统中实现的主要算法包括：若干适于数据流环境的分类与聚类算法：基于信念网络结构模型的VFBN1算法（Learn the structure of a BeliefNet from a very large data set using sampling），VFBN2算法（Learn the structure of a BeliefNet from a very large data set using sampling and a new search procedure），基于决策树模型的VFDT算法（Learn a decision tree from a high-speed data stream or very large data set），基于EM聚类模型的VFEM算法（Performs EM clustering），VFKM（Perform k-means clustering accelerated with sampling）与处理数据流中概念漂移问题的分类算法CVFDT。由于开发者提供了详细的使用说明文档（详见文[1]），此节只简单介绍下笔者将其转换在Windows平台下的若干分类算法的C++版本的使用说明，包括VFDTc算法*与CVFDT算法。所移植的两大分类算法与数据流分类算法实验工具包ETDS的开发平台一致，软件配置与运行环境与ETDS也相同。以下将就VFDTc算法与CVFDT算法的界面操作简单介绍如下：

11.1.1 VFDTc 算法

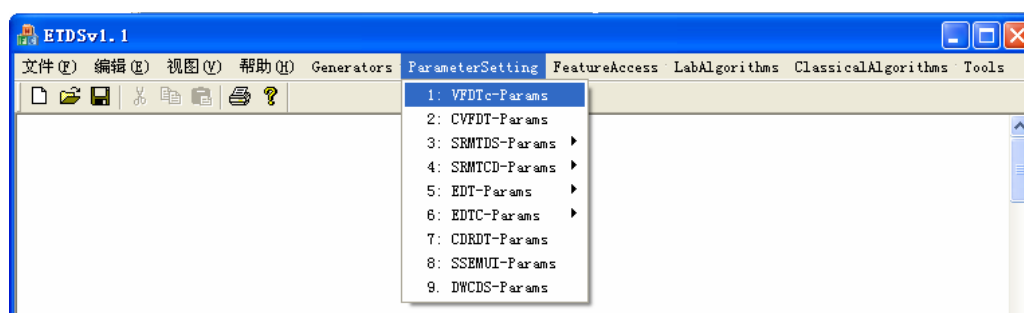


图 11.1 VFDTc 算法参数设定主菜单

VFDTc 算法的运行分为三大步骤，步骤 1：相关参数的设置；步骤 2：数据库特征文

*VFDT 的改进算法，原始 VFML 实验平台封装的是 VFDT 算法，由于此算法仅能处理离散数据流，笔者的实验平台封装的是其改进版本 VFDTc，可以处理混合型数据流。

件的读取；步骤 3：算法运行。以下将针对界面操作逐步介绍算法运行的三大步骤。

11.1.1.1 VFDTc 算法参数设定菜单

由图 11.1 可知，选择 VFDTc 算法菜单弹出相应的参数对话框，如图 11.2 所示。

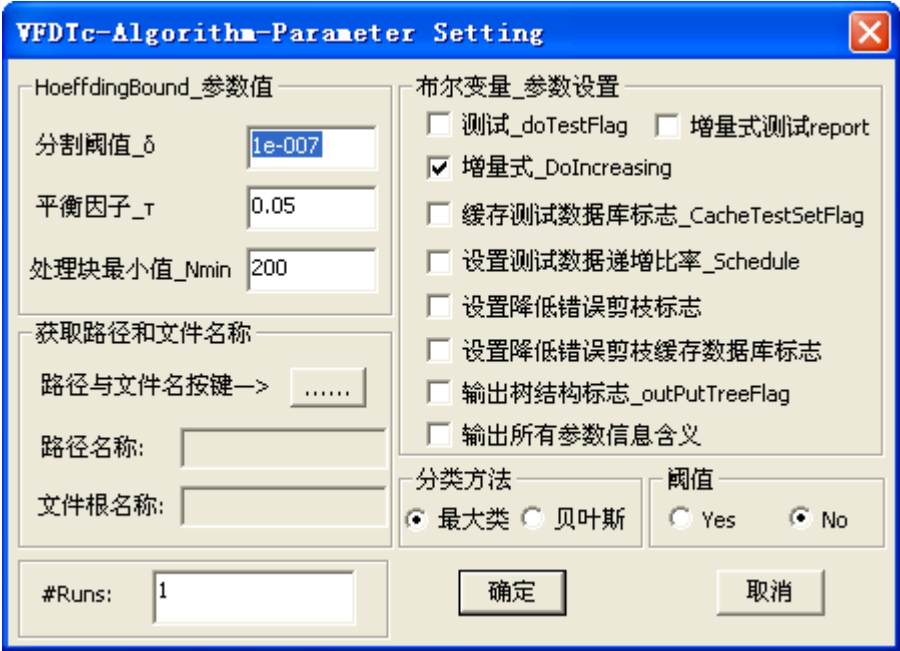


图 11.2 选择 VFDTc-Params 菜单弹出的对话框

现将参数列举如下，见表 11.1。


表 11.1 VFDTc 算法一般参数描述

参数名称	参数意义	类型	界面对应位置	参数值(初始值)
m_fSplitConfidence	Hoeffding Bounds 不等式中的分割阈值	double	分割阈值_δ	10 ⁻⁷
m_fTieConfidence	避免 ties 的阈值	double	平衡因子_τ	0.05
m_iMinChunk	决策树结点中运行一次分割所需的数据量大小	int	处理块最小值_N _{min}	200
.....	设置生成数据库的路径，读取.test 文件	string	路径与文件名按键->	无默认值
m_ShowSourceDirectoryInEdit	只读项，显示读取的文件名称路径	string	路径名称	
m_ShowFileStemInEdit	只读项，显示读取的.test 文件名称的无后缀名称，用来指示.data 文件名与此文件根名一致	string	文件根名称	
m_iRuns	指示当前算法运行次数	int	#Runs	1
m_bDoIncrementalReport	增量式输出分类结果标志	bool	选择框增量式测试 report	false, 即不增量式输出训练中的分类结果
m_bDoIncre	指示当前算法是批量	bool	增量式_DoIncreasing	true; true 表示为

	式处理算法还是增量式处理算法			增量式算法，false 表示为批处理算法
m_bDoTestFlag	/	bool	选择框测试 doTestFlag	false
m_bCacheTestSet	/	bool	缓存测试数据库标志 _CacheTestSetFlag	false，即不设置缓冲测试区标志
m_bUseSchedule	/	bool	设置测试数据递增比率_Schedule	false
m_bREPrune	/	bool	设置降低错误剪枝标志	false
m_bCachePruneSet	/	bool	设置降低错误剪枝缓存数据库标志	false
m_bOutPutTrees	/	bool	输出树结构标志 m_outPutTreeFlag	false
m_bOutPutHelp	/	bool	输出所有参数信息含义	false
m_bBayesMethod	分类方法	bool	最大类 贝叶斯	false，即采用最大类方法
m_bConValuesThres	阈值	bool	Yes No	false，一般在选择采用贝叶斯分类方法时，此参数才设置为有效状态

注：m_bCacheTestSet，m_bUseSchedule，m_bREPrune，m_bCachePruneSet，m_bOutPutTrees，m_bOutPutHelp 等变量主要针对经典 VFDTc 算法，在此处没有特别用处。m_bDoIncrementalReport、m_bConValuesThres 值为 true 时，即处于有效状态时，会相应弹出图 11.3 与图 11.4 的对话框，相应设置增量式输出的事例间隔参数 m_increReportCount（默认值为 10000）与连续属性处理时累积的连续属性值个数阈值 m_maxNumOfConValues（默认值为 1000）。

假设 VFDTc 算法此时采用 Hoeffding Bounds 不等式作为决策树增长结点分割

方法见图 11.2 界面，在此界面中，选择  按钮，打开***.test 文件（以 adult.test）为例，故以下的操作均是在以选择 adult.test 文件之后的界面交互操作。

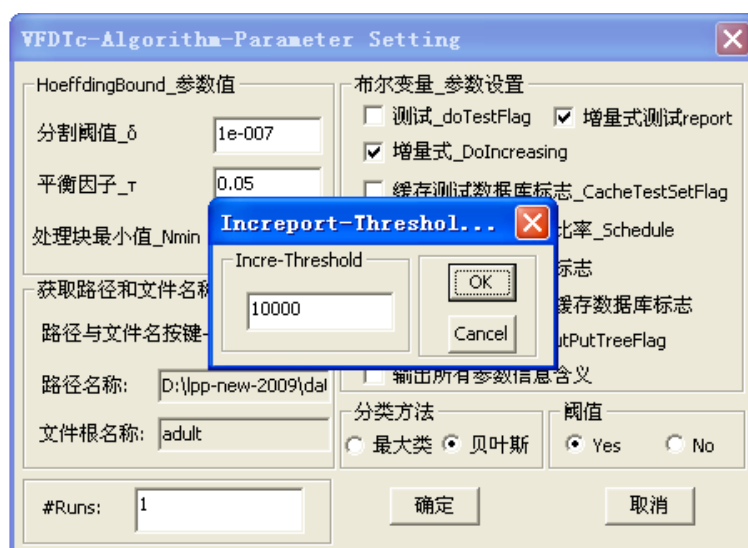


图 11.3 选择增量式测试 report 弹出的对话框

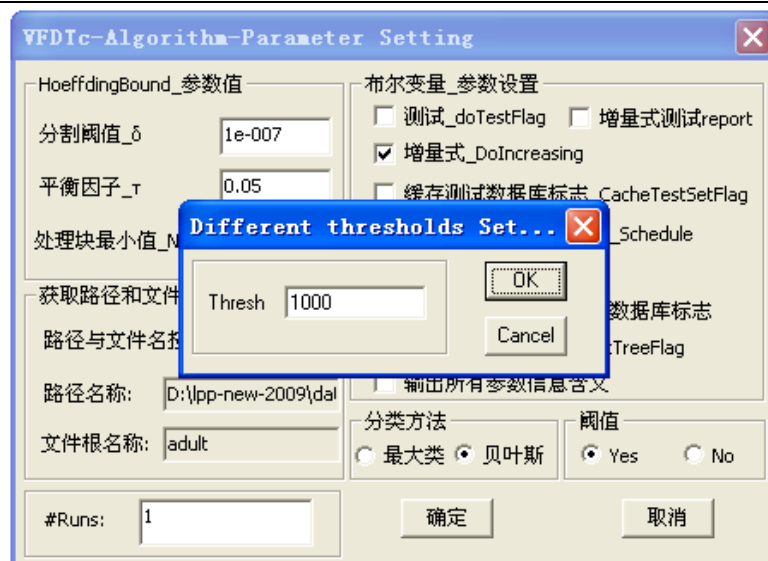
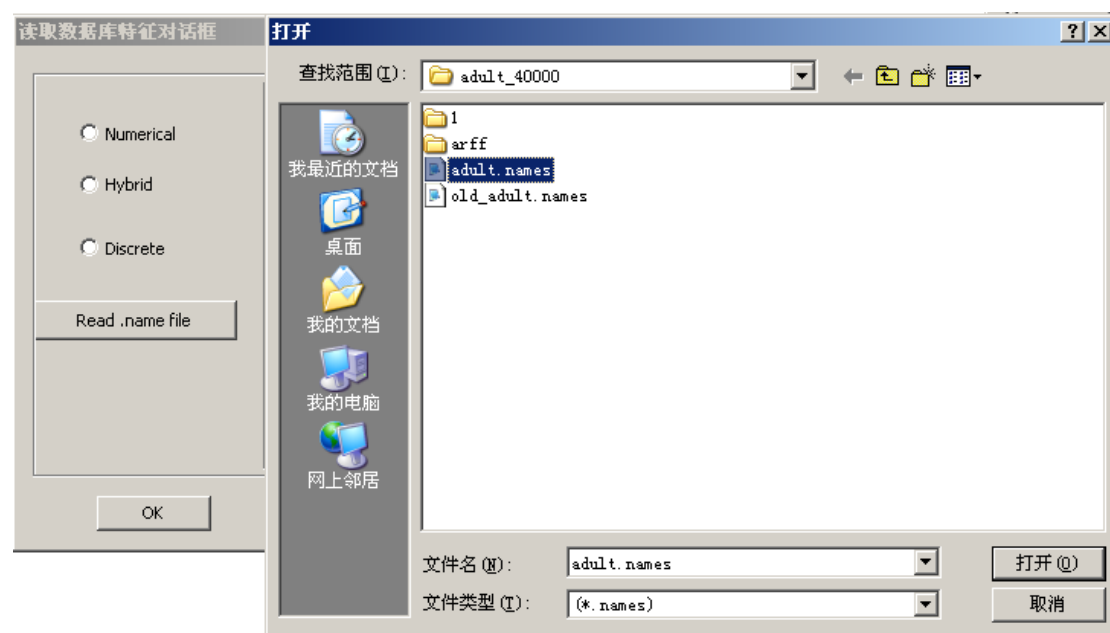


图 11.4 选择贝叶斯分类方法与阈值设定后弹出的对话框

11.1.1.2 VFDtc 算法特征数据库读取与算法运行菜单

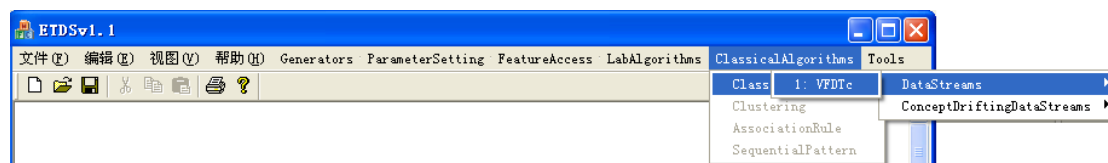
以上涉及的参数设置是算法运行的第一步骤，完成参数设置之后，需要进行以下两步骤得以运行相应的算法，即步骤二：读取待处理的数据库的特征（读取 adult.names 文件），见图 11.5-(a)，图 11.5-(b)；步骤三：完成参数设置与特征读取后，选择对应的算法菜单如 VFDtc 算法，进入选择读取.data 文件的界面，见图 11.5-(c)与图 11.5-(d)，以选择 adult.data 为例，读取了训练集后，选择“打开”按钮，即进入运行 SRMTDS 算法阶段。在算法运行阶段，界面无任何提示状态，一旦算法运行完毕，相应的分类结果在界面显示，如图 11.5-(e)。



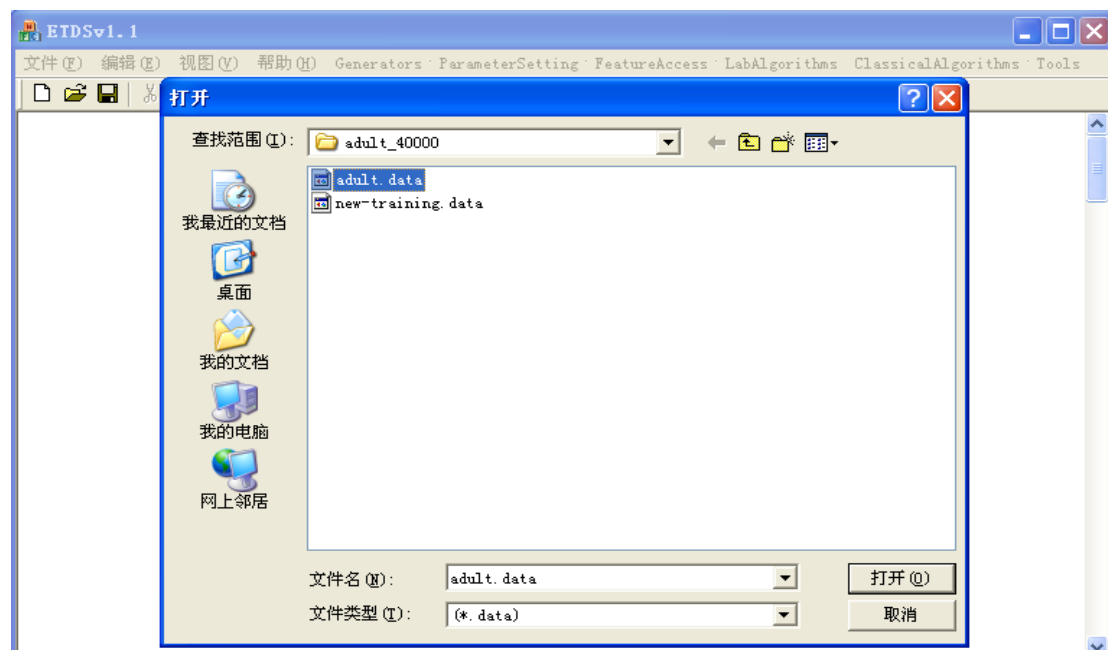
(a) 步骤 2-1—选择 Read.names file 按钮，弹出右侧文件打开框



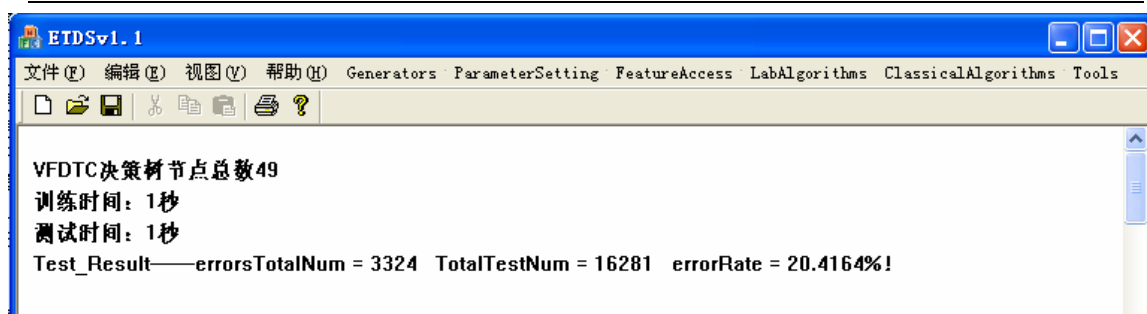
(b) 步骤 2-2—选择步骤 2-1 右侧打开窗口的打开按钮，读取数据库特征并显示



(c)步骤 3-1—选择 VFDTc 算法，弹出打开对话框（如下图）



(d)步骤 3-2—打开训练数据集 adult.data



(e) VFDTc 算法运行完毕时的分类结果（以 adult 数据库为例）

图 11.5 VFDTc 算法运行中的特征读取、算法运行以及结果显示

如果不选择 Naïve Bayes 分类器，选择的是最大类方法，见图 11.6，按下确定后，按照图 11.5 涉及的其他步骤，最后分类算法的执行结果见图 11.7。

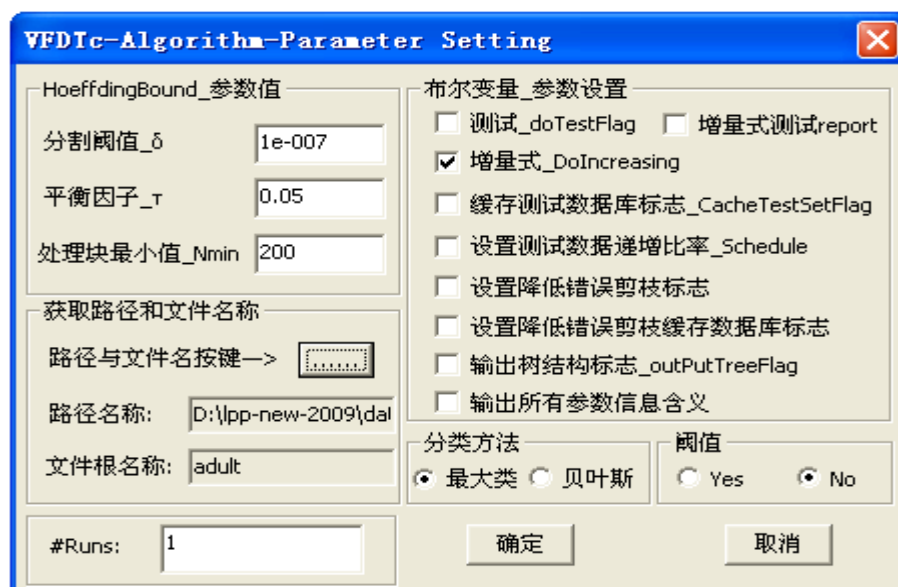


图 11.6 VFDTc 算法采用最大类分类方法的参数设置

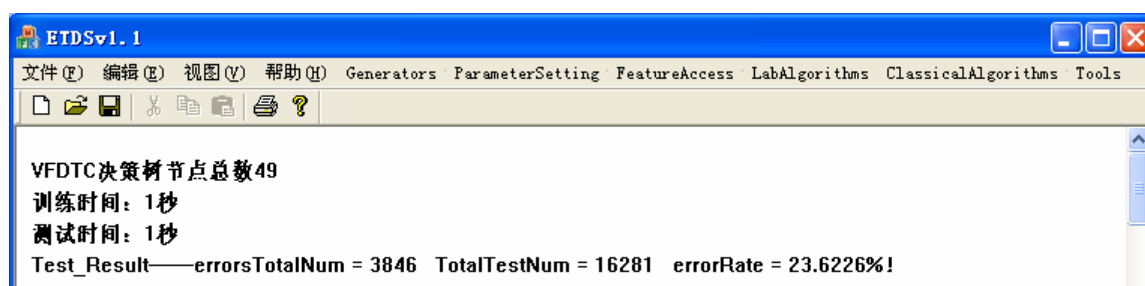


图 11.7 VFDTc 算法采用最大类分类方法的实验结果

11.1.2 CVFDT 算法

与 VFDTc 算法运行步骤相似，CVFDT 算法的运行也分为三大步骤，步骤 1：相关参数

的设置；步骤 2：数据库特征文件的读取；步骤 3：算法运行。以下将针对界面操作逐步介绍算法运行的三大步骤。

11.1.2.1 CVFDT 算法参数设定菜单

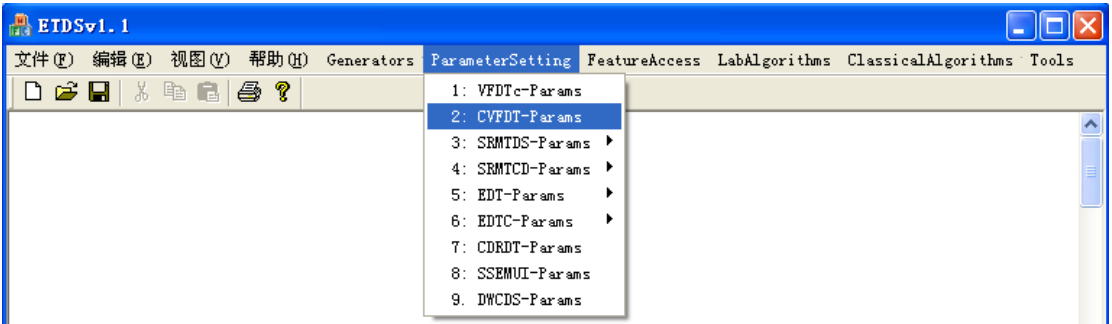


图 11.8 CVFDT 算法参数设定主菜单

由图 11.8 可知，选中 CVFDT 算法菜单，弹出的对话框如图 11.9 所示，若干默认参数已经初始化。



图 11.9 选择 CVFDT 算法菜单弹出参数对话框

现将参数列举如下，见表 11.2。

表 11.2 CVFDT 算法一般参数描述

参数名称	参数意义	类型	界面对应位置	参数值（初始值）
m_fCVSplitThresh	Hoeffding Bounds 不等式中的分割阈值	double	分割阈值_ δ	0.0001
m_fCVTieThresh	避免 ties 的阈值	double	平衡因子_ τ	0.05
m_iCVCheckMinNum	决策树结点中运行一次分割所需的数据量大小	int	处理块最小值_ N_{min}	300
~	设置生成数据库的路径，读取.test 文件	string	路径与文件名按键—>	无默认值
m_strCVPathText	只读项，显示读取的文件名称路径	string	路径名称	
m_strCVFileStem	只读项，显示读取的.test 文	string	文件根名称	

	件名称的无后缀名称, 用来指示.data 文件名与此文件根名一致			
m_bBayesMethod	分类方法	bool	最大类 贝叶斯	false, 即采用最大类方法
m_iCVCheckSize	检测替换子树是否更适宜当前窗口数据时的检测周期	int	测试例子大小初始值 CheckSize	10000
m_iCVWindowSize	滑动窗口大小	int	设置窗口初始值 WindowSize	50000
m_iCVCacheSize	/	int	缓存区大小初始值 CacheSize	10000
m_iCVAltTestNum	判断替换子树是否更适宜当前窗口数据时需要的数据集大小阈值	int	替换分支测试数初始值 AltTestNum	1000
m_fCVSheduleMulti	/	double	测试数据测试比例初始值 SchedulMult	1.44
m_iCVGrowMessg	/	int	总空间消耗初始值 GrowMessage	1000
m_iCVShedulCount		int	测试初始值 ShedulCount	10000
m_CVDoTest	/		doTest	
m_CVDoIncrReport	增量式输出分类结果标志	bool	IncrReport	false; true 表示为增量式算法, false 表示为批处理算法
m_bCVfinalOutput	/		finOutput	false;
m_iRuns	指示当前算法运行次数	int	#Runs	1

注: m_iCVCacheSize, m_fCVSheduleMulti, m_iCVGrowMessg, m_iCVShedulCount, m_CVDoTest 等变量主要针对经典 CVFDT 算法设定, 在此处没有特别用处。m_CVDoIncrReport 值为 true 时, 即处于有效状态时, 会相应弹出图 11.10 的对话框, 相应设置增量式输出的事例间隔参数 m_incrReportCount (默认值为 10000)。

如图 11.9, 在此界面中, 选择 **浏览文件** **>>>** 按钮, 打开***.test 文件(以 adult.test)为例, 故以下的操作均是在以选择 adult.test 文件之后的界面交互操作。

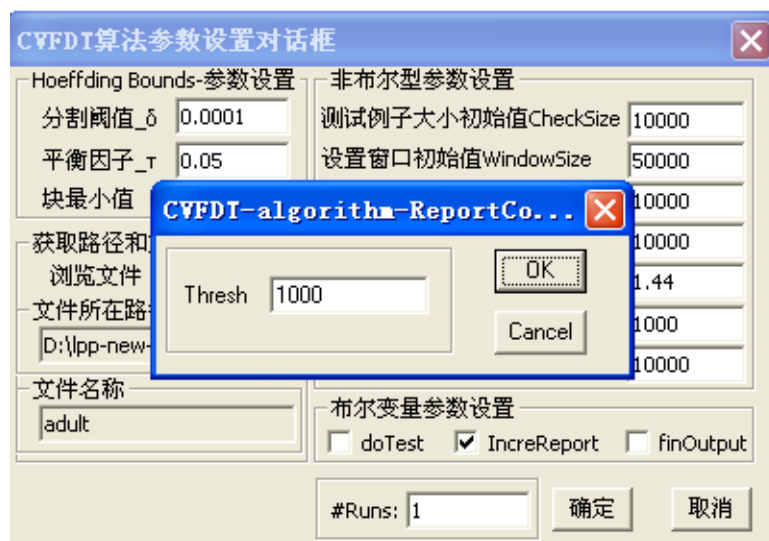
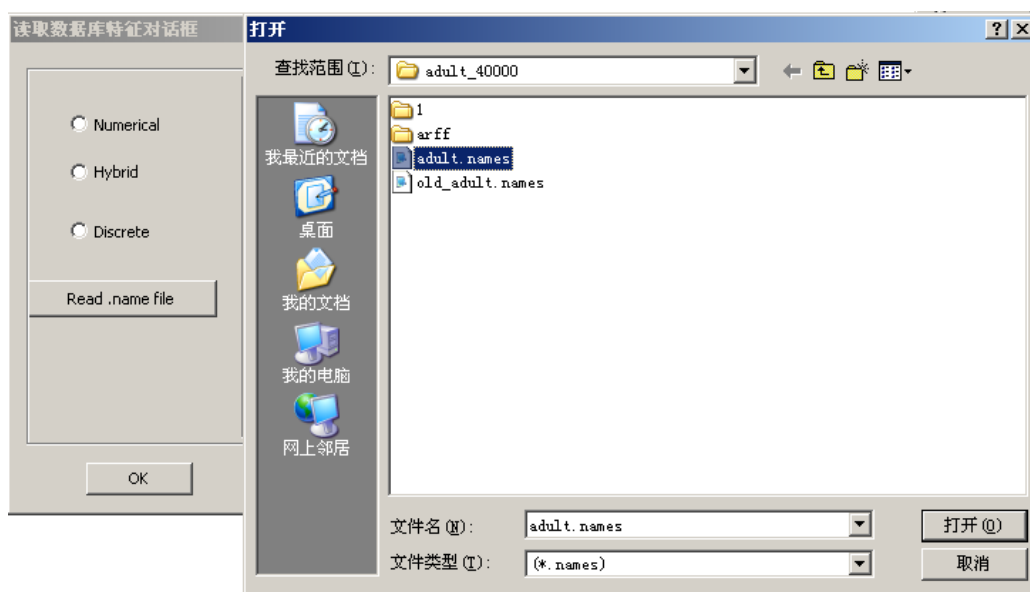


图 11.10 勾选 IncrReport 选择框时弹出的对话框

11.1.2.2 CVFDT 算法特征数据库读取与算法运行菜单

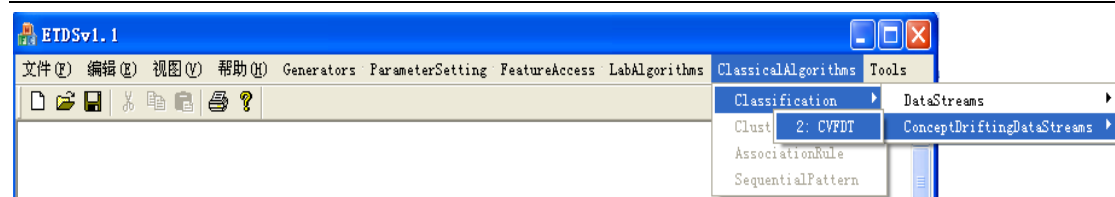
以上涉及的参数设置是算法运行的第一步骤，完成参数设置之后，需要进行以下两步骤得以运行相应的算法，即步骤二：读取待处理的数据库的特征（读取 `adult.names` 文件），见图 11.11-(a)，图 11.11-(b)；步骤三：完成参数设置与特征读取后，选择对应的算法菜单如 CVFDT 算法，进入选择读取 `.data` 文件的界面，见图 11.11-(c) 与图 11.11-(d)，以选择 `adult.data` 为例，读取了训练集后，选择“打开”按钮，即进入运行 CVFDT 算法阶段。在算法运行阶段，界面无任何提示状态，一旦算法运行完毕，相应的分类结果在界面显示，如图 11.11-(e)。



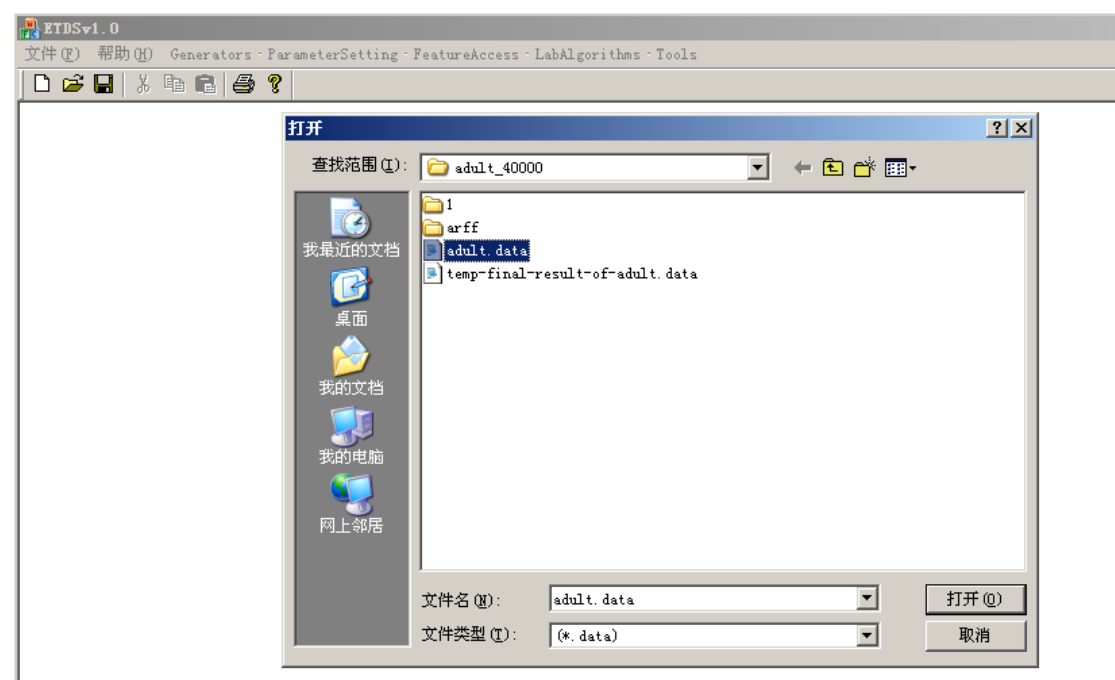
(a) 步骤 2-1—选择 Read.names file 按钮，弹出右侧文件打开框



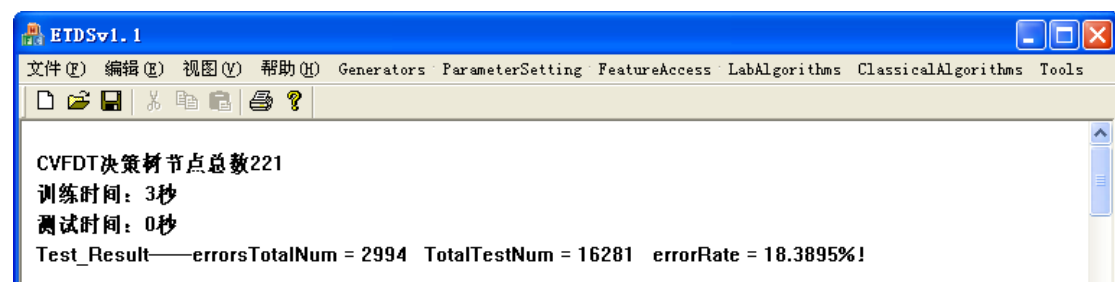
(b) 步骤 2-2—选择步骤 2-1 右侧打开窗口的打开按钮，读取数据库特征并显示



(c) 步骤 3-1—选择 CVFDT 算法，弹出打开对话框（如下图）



(d) 步骤 3-2—打开训练数据集 adult.data



(e) CVFDT 算法运行完毕时的分类结果(以 adult 数据库为例)

图 11.11 CVFDT 算法运行中的特征读取、算法运行以及结果显示

11.2 MOA

MOA (Massive Online Analysis) 是一个用于在线数据流学习的开源环境^[2]。在 WEKA (the Waikato Environment for Knowledge Analysis) 开发团队设计的一种适于数据流环境的实验工具，用 java 编程实现。MOA 提供开源代码与界面操作，MOA 包含了在线与离线方法以及评估工具，实现了 boosting, bagging 和 Hoeffding Tree 三种方法，所有这些方法均可在叶