一种基于基本显露模式的分类算法*)

范 明 刘孟旭 赵红领
(郑州大学计算机科学系 郑州450052)

摘 要 本文提出了一种新的基于EP的分类法CEEP。CEEP仅使用最短的EP(eEP)建立分类器,并使用不同于早先的基于EP的分类法(如,CAEP)的评分标准。文中还讨论了eEP的有效挖掘,最小支持度和最小增长率阈值的自适应选取等问题。在UCI机器学习库中的12个数据集上的实验表明,本文的分类方法具有很好的分类正确率。如何保证eEP有足够的覆盖率,以及如何处理稀有类的分类,尚待进一步研究。此外,如何将装袋(bagging)和推进(bootstrap)的思想与CEEP的方法相结合,进一步提高分类的正确率,也是值得深入研究的问题。

关键词 数据挖掘,分类,显露模式

Classification by Essential Emerging Patterns

FAN Ming LIU Meng-Xu ZHAO Hong-Ling
(Department of Computer Science, Zhengzhou University, Zhengzhou 450052)

Abstract Emerging patterns (EPs) are itemsets whose supports change significantly from one data class to another. It has been shown that they are useful for constructing accurate classifiers. However, the existing EP-based classifiers may suffer from two major deficiencies: (1) they use a large number of EPs, which may lead to high processing overhead; and (2) their scoring method based on growth rate and support may reduce the contribution of EPs with high differentiating power and low support, which may lead to misclassification. This work proposes a novel classification method, CEEP (Classification by Essential Emerging Patterns), which uses a special kind of EPs, called essential Emerging Patterns (eEPs), and a growth-rate-based scoring method to construct classifiers. Mining eEPs is much easer than mining EPs, and using eEPs only is sufficient to construct accurate classifiers. Our experiment study carried on 12 benchmark datasets from the UCI Machine Learning Repository shows that CEEP performs comparably with other state-of-the-art classification methods such as NB, C5.0, CBA, CMAR, CAEP and BCEP in terms of overall predictive accuracy.

Keywords Data mining, Classification, Emerging pattern

1 引言

分类是重要数据分析任务之一,在商业、金融、电讯、DNA分析、科学研究等诸多领域具有广泛的应用。统计学、机器学习、神经网络等领域的研究者提出了很多分类方法,大部分算法是内存驻留算法,适用于小型数据集。随着数据集的数据量和数据维数的增加,建立高效的、适用于大型数据集的分类算法已经成为数据挖掘的一个挑战性问题。

基于显露模式(Emerging Pattern, EP)的分类方法是针对大型数据集的分类提出的。EP是G. Dong和J. Li提出的一种新的知识模式,这些模式捕获目标类和非目标类上多组属性之间的不同,具有很好的分类性能[2]。基于EP的分类法的核心问题是:使用什么样的EP,如何有效地从大型数据集中挖掘这些EP,以及如何使用这些EP(建立评分标准)确定新样本的所属类。

第一个基于EP的分类算法是G. Dong等提出的CAEP算法[3]。CAEP同时使用增长率和支持度来度量每个EP在指导分类中的贡献,并聚合每个EP的贡献来计算待分类样本属于每个类的得分,从而确定样本的所属类。由于EP的数量太大,为避免使用所有EP,CAEP采用了一些归约策略以压缩分类时所使用的EP的数量,但用于分类的EP数量仍然很大。Li, Dong和Ramamohanarao提出了利用JEP分类的算法JEP-Classifier[6]。JEP具有很好的区分能力,但是如果仅使用JEP常常没有足够的覆盖率,迫使过多的实例不得不采用多数表决的方法决定其所属的类。之后提出的一些基于EP的分类方法,如文[7,11],采用了不同的归约策略以减少EP/JEP的数量,但一直沿用CAEP的评分策略。最近,Fan和Ramamohanarao提出使用基本显露模式(essential Emerging Pattern, eEP)建立基于EP的Bayes分类方法BCEP,取得了很好的分类效果[4]。BCEP是一种"懒散"的分类法,分类速度相对较慢。

本文提出一种新的基于EP的分类算法CEEP(Classification by Essential Emerging Patterns)。CEEP使用eEP建立分类器。事实上,eEP是最短的EP,并且具有一定的支持度。尽管挖掘所有EP是困难的,但eEP可以有效地挖掘。与已有的基于EP的分类法(如CAEP)不同,CEEP采用了更合理的、基于EP增长率的评分标准,用于确定未知样本所属的类。在UCI机器学习库的12个数据集上的测试表明,CEEP的分类正确率优于CAEP,并且可以与已知的最好的分类法媲美。

*)本文的工作得到河南省自然科学基金(项目号:0211050100)的资助。

本文第2节简要介绍EP的基本概念。第3节介绍CEEP分类法的主要思想和基本成分，包括eEP的有效挖掘，评分标准，最小支持度和最小增长率阈值的自适应选取。第4节报告我们在UCI机器学习库的12个数据集上实验结果，并与NB，C5.0，CBA，CMAR，CAEP和BCEP等分类法进行比较。

2 基本概念

设训练数据集是一个标准的关系表，它包含 N 个实例(样本)，每个实例有 m 个不同的属性。所有的连续属性已经离散化，并且映射到一个连续的正整数的集合中。假定 N 个实例被划分为 K 个已知类 $C_1, C_2, \cdots, C_K$，并给定了每个训练实例所属的类。项是属性名和属性值的对偶。令 $I = \{i_1, i_2, \cdots, i_m\}$ 是样本中出现的项的全集。I 的子集 $X \subseteq I$ 称作项集。使用项的概念，每个样本都是一个包含 m 个元素的项集，而训练数据集是项集的集合。

定义1 项集 X 在数据集 D 上的支持度，记为 $sup_D(X)$，定义为 $sup_D(X) = count_D(X)/|D|$，其中 $count_D(X)$ 是数据集 D 中包含 X 的样本个数，而 $|D|$ 是数据集 D 中样本的总数。

定义2 给定两个不同的数据集 D' 和 D，项集 X 从 D' 到 D 的增长率 $GR(X, D', D)$ 定义如下：

$$GR(X,D',D)=\begin{cases}0 & \text{如果 } sup_D(X)=sup_{D'}(X)=0\\ \infty & \text{如果 } sup_D(X)\neq 0, sup_{D'}(X)=0\\ sup_D(X)/sup_{D'}(X) & \text{其他}\end{cases}$$

如果数据集 D' 和 D 分别是 C' 类和 C 类样本的集合，则增长率是项集 X 从类 C' 到类 C 支持度变化显著程度的度量。

定义3 给定增长率阈值 $\rho > 1$，如果项集 X 是从 D' 到 D 的增长率 $GR(X, D', D) \geq \rho$，则称 X 是从 D' 到 D 的 ρ-EP(或EP)，或者简单地称 X 是 D 的EP。如果 X 的增长率为 ∞，则称 X 为 D 的JEP(Jumping EP)。

EP是那些从一个数据集到另一个数据集支持度发生很大变化的项集，这些项集能够很好地捕获目标类和非目标类上多个属性之间的不同，所以具有很好的区分性。

例1 考虑Mushroom数据集，设

X = {(BRUISES = no), (GILL_SPACING = close), (VEIL_COLOR = white)}

Y = {(ODOR = none), (GILL_SIZE = broad), (RING_NUMBER = one)}

我们有

项集	支持度	
	有毒类	可食类
X	81.4%	3.8%
Y	0%	63.9%

项集 X 在可食类中的支持度为3.8%，在有毒类中的支持度为81.4%；作为有毒类的EP，X 的增长率为0.814/0.038 = 21.4。如果样本 S 包含模式 X，就可以以很高的概率认为它属于有毒类。项集 Y 在有毒类中的支持度为0%，在可食类中的支持度为63.9%；作为可食类的EP，Y 的增长率为 ∞(Y 是可食类的JEP)。如果样本 S 包含模式 Y，我们几乎可以断定它属于可食类。

EP具有很好的区分能力。然而，待分类的数据集可能存在大量EP(数以万计)，并且这些EP不是独立的(存在包含关系)。实践表明使用所有的EP进行分类并非很有效。H. Fan和K. Ramamohanarao提议使用一类特殊的EP，称为eEP，建立基于EP的Bayes分类器[4]。eEP定义如下：

定义4 项集 X 称为 D 上的eEP，如果 X 是 D 的EP，并且 X 在 D 中的支持度不小于预先指定的最小支持度阈值 ξ，而 X 的任何真子集都不满足上述条件。

事实上，eEP是那些"最短的"EP。由于如下原因，eEP被认为是EP中最具表达能力的模式：

- 很高甚至无穷大的增长率确保eEP具有很好的区分能力。
- 最小支持度阈值确保每个eEP至少覆盖一定数量的样本，从而具有一定的统计意义，确保它们的实用性。
- eEP的超集对于分类并没有多大用处，其理由如下：假定 $E_1 \subset E_2$，并且 E_1 是eEP。根据定义，E_1 具有很高的增长率，并且被 E_2 覆盖的样本一定都被 E_1 覆盖，从而 E_2 并不提供比 E_1 更多的分类信息。

3 CEEP分类法

现在，我们介绍CEEP分类法的主要思想和基本成分：(1)如何挖掘每个类的eEP；(2)单个eEP对确定类成员关系的贡献，以及如何聚集每个eEP的贡献，确定未知样本的类；(3)如何确定最小支持度阈值 ξ 和最小增长率阈值 ρ。

3.1 挖掘eEP

对于给定的训练数据集 DB，它包含 $C_1, C_2, \cdots, C_K$ 个类。为建立CEEP分类器，我们需要对于 $i = 1, 2, \cdots, K$，某支持度阈值 ξ 和增长率阈值 ρ，求 C_i 类和非 C_i 类的eEP。

文[2]的研究表明：给定最小支持度阈值和最小增长率阈值，EPs可以用边界表示。设 L 和 R 是项集的集合，边界 $\langle L, R\rangle$ 表示的项集的集合为：

$$[L,R] = \{Y \mid \exists X \in L, \exists Z \in R, X \subseteq Y \subseteq Z\}$$

例如，边界 $\langle\{\{a\},\{b,c\}\},\{\{a,b,c,d\}\}\rangle$ 代表这样的一些项集：它是 $\{a\}$ 的超集，并且是 $\{a,b,c,d\}$ 的子集；或者它是 $\{b,c\}$ 的超集，并且是 $\{a,b,c,d\}$ 的子集。事实上，eEPs正是EPs边界表示的左边界。给定两个数据集 D_1 和 D_2，最小支持度阈值和最小增长率阈值，利用文[2]的边界算法，可以计算 D_1 到 D_2 的所有eEP。

然而，挖掘eEP还有更快的算法。我们开发了一种基于模式树(P-树)的挖掘eEP的有效算法。与FP-树[5]类似，P-树存储了训练数据集中所有的项信息。不同的是，P-树维护了类信息，以支持挖掘eEP。该算法采取了模式增长的挖掘方法，并直接地在P-树中挖掘所有的eEP，而不需要附加的空间。该算法的思想部分地与文[12]类似。限于篇幅，我们将另文详细讨论。

3.2 使用eEP分类

为确定待分类样本 S 所属的类，C_i 类的每个eEP都试图确定 S 是否属于 C_i 类。设 D_i 是 C_i 类训练样本的集合，D_i' 是非 C_i 类训练样本的集合，X 是 C_i 类的eEP。如果 X 不在 S 中出现，则 X 不能为确定 S 是否属于 C_i 类做出判断。如果 X 在 S 中出现，由于 X 在 C_i 类出现的频率(支持度)是在非 C_i 类出现的频率的 $GR(X, D_i', D_i)$ 倍，因此 X 将以几率 $\frac{GR(X,D_i',D_i)}{GR(X,D_i',D_i)+1}$ 判定 S 属于类，而以几率 $\frac{1}{GR(x,D_i',D_i)+1}$ 判定 S 不属于类。如果 X 是 C_i 类的JEP，则 $GR(X, D_i', D_i) = \infty$。此时，我们令 $\frac{GR(X,D_i',D_i)}{GR(X,D_i',D_i)+1}=1$，$\frac{1}{GR(X,D_i',D_i)+1}=$

0。

非 C_i 类的 eEP 对于确定 S 是否属于 C_i 类也有贡献。令 Y 是非 C_i 类的 eEP，它在 S 中出现。如果 Y 的增长率很大，Y 对确定 S 属于 C_i 类的影响可以忽略。然而，当 Y 的增长率不太大(如，$GR(Y,D_i,D'_i)<5$)时，Y 对确定 S 属于 C_i 类的影响相当大。一般地，我们取 Y 确定 S 属于 C_i 类的几率为 $\frac{1}{GR(Y,D_i,D'_i)+1}$。

为了对样本 S 进行分类，我们需要组合 C_i 类和非 C_i 类的每个 eEP 的贡献，计算 S 属于 C_i 类的得分 $score(S,C_i)$。对于 $i=1,2,\cdots,K$，令 $PS(S,C_i)=\{X|X$ 是 D_i 的 eEP，并且 X 在 S 中出现$\}$，$NS(S,C_i)=\{X \mid X$ 是 D'_i 的 eEP，并且 X 在 S 中出现$\}$。S 属于 C_i 类的得分 $score(S,C_i)$ 用下式计算：

$$score(S,C_i)=\sum_{X\in PS(S,C_i)}\frac{GR(X,D'_i,D_i)}{GR(X,D'_i,D_i)+1}+\sum_{Y\in NS(S,C_i)}\frac{1}{GR(Y,D_i,D'_i)+1} \quad (1)$$

CEEP 分类法将使用如下规则对 S 进行分类：将 S 划归得分最高的类。如果得分最高的类不唯一，则将 S 划归得分最高的多数类。

注意：CAEP 使用下式计算 S 属于 C_i 类的得分 $score(S,C_i)$[3]

$$score(S,C_i)=\sum_{X\in S(S,C_i)}\frac{GR(X,D'_i,D_i)}{(X,D'_i,D_i)+1}\times sup_{D_i}(X) \quad (2)$$

其中，$S(S,C_i)=\{X|X$ 是 D_i 的 EP，并且 X 在 S 中出现$\}$。比较(1)式和(2)式，除使用 eEPs 和使用 EPs 的差别之外，二者的不同之处有两点。

第一点不同是：CAEP 仅考虑 C_i 类 EP 的贡献，而 CEEP 在计算 C_i 类的得分时不仅考虑 C_i 类 eEP 的贡献，而且考虑非 C_i 类 eEP 的贡献。当非 C_i 类 eEP 的增长率很高时(例如，大于20)，非 C_i 类 eEP 影响可以忽略。然而，当非 C_i 类 eEP 的增长率不是很高时(例如，小于10)，这种影响并不能忽略。在我们的实验中，12个数据集中的9个的最小增长率阈值都小于10，忽略非 C_i 类 eEP 的影响显然是不合理的。

第二点不同是：在组合每个 EP 的贡献时，CAEP 用 X 在 C_i 类的支持度 $sup_{D_i}(X)$ 加权，而 CEEP 用对每个 X 取相同的权值1。事实上，X 在 C_i 类的支持度反映的是 X 可以对多大比例的样本分类起作用。当 X_1 和 X_2 都是 EP，并在给定的待分类样本 S 中出现时，X_1 和 X_2 对确定 S 所属类的贡献的权重应当是一样的；即，它们的贡献仅由它们的区分能力(增长率)确定。考虑下面的例子：

例2 假定训练数据集包含2000个样本，两个类的大小一样。C_1 类和 C_2 类各有1000个样本。E_1 覆盖了900个 C_1 类样本，450个 C_2 类样本。E_2 覆盖了1个 C_1 类样本，300个 C_2 类样本。显然，E_1 是一个 C_1 类的 EP，它的增长率是2。E_2 是一个 C_2 类的 EP，增长率是300。

考虑样本 S，它仅包含两个 eEP，E_1 和 E_2。因为 C_1 类和 C_2 类中都存在一定数量的包含 E_1 的实例样本，所以依据 E_1 我们不能断定 S 属于哪一个类。然而 C_1 类中几乎不存在包含 E_2 的样本，而 C_2 类中有大量的样本包含 E_2，所以几乎可以断定 S 不属于 C_1 类。综合考虑 E_1 和 E_2，S 显然应该属于 C_2 类。使用 CAEP 的评分方法，S 将被划归 C_1 类，而使用 CEEP 的评分方法，S 将被划归 C_2 类。该例表明 CEEP 的评分标准比 CAEP 更合理。我们的实验结果进一步支持了这一断言。

3.3 最小支持度和最小增长率阈值的确定

影响 CEEP 分类正确率的主要因素是 eEP 的区分度和覆盖率，而它们主要取决于两个相关的参数：最小支持度和最小增长率阈值。一般地，固定最小支持度阈值，较高的增长率导致产生具有较高区分度的 eEP(有利于建立更好的分类器)，但可能降低覆盖率(不利于建立好的分类器)。而固定增长率，支持度越高，覆盖率就越低；但支持度过低，所产生的 eEP 不具有一般性，也不利于建立好的分类器。

我们的实验表明，对于大多数数据集，最小支持度阈值的较好选择大约在1%左右。但是，最小增长率阈值的较好选择却因数据集不同而相差悬殊。例如，当最小支持度阈值为1%时，对于数据集 Mushroom，Australian，German，Cleve 和 Hearts，最小增长率阈值分别取35，4.4，3.1，2.1和1.8时 CEEP 分类的精度最高。因此，给定训练数据集，我们需要在训练阶段确定两个参数：最小支持度阈值 ξ 和最小增长率阈值 ρ。

对于给定的数据集 DB，CEEP 采取随机抽样，将 DB 分成10个大致相等的数据集 $DB_1,DB_2,\cdots,DB_{10}$。交替地固定最小支持度阈值 ξ 和最小增长率阈值 ρ 中的一个，采用十折交叉验证自适应地选择另一个，直到分类正确率不再提高或满足要求为止。例如，固定最小支持度阈值 $\xi=0.01$，CEEP 通过如下步骤选择最小增长率阈值：

(1)置初值：选取最小增长率阈值，例如 $\rho=2$；

(2)使用 ξ 和 ρ 的当前值，对于 $k=1,2,\cdots,10$

(a)令 $D=\bigcup_{i\neq k}DB_i$，在 D 中挖掘 C_i 类和非 C_i 类的 eEP($i=1,\cdots,K$)；

(b)对于 DB_k 中每个样本 S

使用(1)式计算 S 属于 C_i 类的得分 $score(S,C_i)$($i=1,\cdots,K$)，并按3.2节的分类规则确定 S 所属类；

(3)统计分类正确率；如果分类正确率不再提高或满足要求，则结束；

(4)将 ρ 增加或减少一个增量，转(2)。

一般地，当分类正确率上升时，可以提高 ρ 值；反之，降低 ρ 值。增量可以取固定值(如0.1)，也可以用递减的增量。最小支持度阈值 ξ 的自适应选择可以类似地进行。这种参数的自适应选择虽然不能保证得到最佳的参数值 ξ 和 ρ，但通常可以得到较好的。

可以对不同的类 C_i 使用不同的最小支持度阈值 ξ_i 和最小增长率阈值 ρ_i。这样需要确定2K 个参数，使得训练时间将大大增加。带来的好处是有助于提高最终分类器的分类正确率。我们的经验表明，对于 $K=2$，可以采用不同的最小支持度阈值 ξ_i 和最小增长率阈值 ρ_i。当 K 较大时，使用不同的最小支持度阈值 ξ_i 和最小增长率阈值 ρ_i 的好处并不明显。

3.4 CEEP 分类器的构造与使用

现在，我们将 CEEP 分类器的构造与使用总结如下：

假设训练数据集 DB 包含 K 个类。在训练阶段，CEEP 的任务是：

(1)使用3.3节的方法确定最小支持度阈值 ξ_i 和最小增长率阈值 ρ_i。

(2)使用最小支持度阈值 ξ_i 和最小增长率阈值 ρ_i，对于 $i=1,\cdots,K$，在训练数据集 DB 上挖掘 C_i 类和非 C_i 类的 eEP。

这些 eEP 与3.2节评分策略和分类规则构成了一个 CEEP 分类器。对于任意待分类样本 S，CEEP 分类器将按以下步骤确定 S 所属的类：

(1)对于 $i=1,\cdots,K$，求 $PS(S,C_i)$ 和 $NS(S,C_i)$；使用

(1)式,计算 S 属于 C_i 类的得分 $score(S,C_i)$;

(2)将 S 划归得分最高的类。如果得分最高的类不唯一,则将 S 划归得分最高的多数类。

4 实验结果及其分析

为了验证CEEP的分类准确性,我们使用UCI机器学习库[1]中的12个数据集作为实验数据集,并将实验结果与NB、C5.0、CBA、CMAR、CAEP以及BCEP比较。

实验环境为900Mhz Pentium III CPU,128MB内存,30GB硬盘,操作系统为Microsoft Windows 2000,编程软件为Microsoft Visual C++。实验采用十折交叉验证的方法来统计分类的准确性。即,将数据集划分成十个互不相交的子集或"折" $DB_1,DB_2,\cdots,DB_{10}$,每个折大小大致相等。训练和检验进行10次。在第 i 次,DB_i 用测试集,其余的子集都用于训练分类法。误分类率估计是十次迭代误分类的样本数之和除以样本总数,而分类的正确率等于1一误分类率。

表1给出了在这12个数据集中各种分类方法的正确率。其中,CEEP的分类正确率是我们的实验结果,其它分类法的结果数据取自相应文献[3,4,7~9]。表中"—"表示缺乏数据,在求平均正确率时取诸分类法的平均值。

表1 分类正确率:NB,C5.0,CBA,CMAR,CAEP,BCEP和CEEP比较

数据集	NB	C5.0	CBA	CMAR	CAEP	BCEP	CEEP
Adult	84.12	85.54	—	—	83.09	85.00	81.85
Australian	85.65	84.93	84.90	86.10	86.21	86.40	87.83
Cleve	82.78	77.16	82.80	82.20	83.25	82.41	84.33
Diabete	75.13	73.03	74.50	75.60	67.30	76.80	74.08
German	74.94	71.90	73.40	74.90	72.50	74.50	73.40
Heart	82.22	76.30	81.90	82.20	83.70	81.85	84.07
Iono	89.45	91.45	92.30	91.50	89.76	93.24	92.57
Mushroom	99.68	100	—	—	98.82	100	99.90
Pima	75.90	75.39	72.90	75.10	75.00	75.66	73.03
Sonar	75.40	76.00	77.50	79.40	78.30	78.40	85.50
Vechile	61.12	69.82	68.70	68.80	66.32	68.5	67.44
Lymph	81.86	78.29	77.80	83.10	74.38	83.13	84.38
平均	80.69	79.98	80.86	81.88	79.89	82.16	82.37

CEEP在12个数据集中的5个获胜,并具有最好的平均正确率。与NB(朴素Bayes分类法)相比,CEEP在9个数据集上胜出。与C5.0(一种决策树分类法)相比,CEEP在8个数据集上胜出。CBA和CMAR是两种基于关联规则的分类法,CEEP分别在10个数据集中的7个和6个上胜出。本质上,基于关联规则的分类法和基于EP的分类法都是使用某种频繁模式进行分类。由于EP是具有很高区分度的频繁模式,采用合理的评分标准,CEEP的分类性能总体上优于基于关联规则的分类法。与早期的基于EP的分类法CAEP相比,CEEP在10个数据集上获胜。这表明CEEP的评分标准比CAEP的标准更合理。与基于EP的Bayes分类法BCEP相比,虽然CEEP的平均正确率略高于BCEP,但是BCEP在更多的数据集上获胜(7比5)。

与NB、C5.0、CAEP和BCEP相比,CEEP在Adult数据集上的性能最差。进一步的分析表明,算法挖掘出的eEP不具有足够高覆盖率,许多样本的分类只能靠多数表决确定。这严重地影响了CEEP的分类准确性。

参考文献

1 Blake C, Merz C. UCI repository of machine learning databases. 1998 [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science

2 Dong G, Li J. Efficient mining of emerging patterns: Discovering trends and differences. In: Proc. of KDD'99, San Diego, USA, Sept. 1999. 15~18

3 Dong G, Zhang X, Wong L, Li J. CAEP: Classification by Aggregating emerging patterns. In: Proc. of the 2nd Int'l Conf. On Discovery Science (DS'99), Tokyo, Japan, Dec. 1999. 30~42

4 Fan H, Ramamohanarao K. Bayesian Approach to use Emerging Patterns for Classification. In: Proc of 14th Australasian Database Conf. Feb. 2003. 39~48

5 Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In: Proc. of the 2000 ACM-SIGMOD Intl. Conf. on Management of Data, May 2000. 1~2

6 Li J, Dong G, Ramamohanarao K. JEP-Classifier: Classification by Aggregating Jumping Emerging Patters. Knowledge and Information Systems, 2001, 3(2): 131~145

7 Li J, Dong G. Ramamohanarao K. Making Use of the Most Expressive Jumping Emerging Patterns for Classification. In: Pro. of 2000 Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD'00), 2000. 220~223

8 Li W, Han J, Pei J. CMAR: Accurate and efficient classification based on multiple class-association rules. In: ICDM'01, San Jose, CA, Nov. 2001. 369~376

9 Liu B, Hsu W, Ma Y. Integrating classification and association rule mining. In: KDD'98, New York, NY, Aug. 1998. 80~86

10 Zheng Z, Webb G I. Lazy learning of Bayesian rules. Machine Learing, 41: 53~84

11 李曼,范明.一种新颖的基于最有效的跳跃显露模式的分类法.计算机科学,2002,29(8.增刊A):73~76

12 范明,王秉政.一种直接在Trans-树中挖掘频繁模式的新算法.计算机科学,2003,30(8):120~127