基于类标号扩展的半监督特征选择算法

王 博[1] 贾 焰[1] 田 李[2]
(国防科技大学计算机学院 长沙 410073)[1] (94326部队 济南 250023)[2]

摘 要 特征选择是数据挖掘、机器学习等领域的重要内容,在缺乏已标记样本的情况下,如何有效选择特征是一个非常值得研究的问题。基于集合间相关度与自相关度的定义,提出了一种新颖的半监督特征选择方法,从原始、少量、且已标记的训练样本出发,通过扩展类标号得到最终的聚类效果,采用复合的评价方法作为衡量特征子集的标准。大量实验结果表明,该算法是有效的。

关键词 特征选择,半监督,集合相关度,集合自相关度

Semi-supervised Feature Selection Algorithm Based on Extension of Label

WANG Bo[1] JIA Yan[1] TIAN Li[2]
(School of Computer Science, National University of Defense and Technology, Changsha 410073, China)[1]
(PLA No. 94326, Jinan 250023, China)[2]

Abstract Feature selection is an important step during data mining and machine learning. With the lack of labeled instances, the problem of effective selection is worthy of consideration. This paper proposed a novel semi-supervised feature selection algorithm based on the definition of inter-set and intra-set correlation, which starts from the original and small labeled samples and gains the final clusters by extension of labels. A complex evaluation was utilized as criterion to find optimal feature subset. Finally, the experimental results demonstrate the efficacy of the algorithm.

Keywords Feature selection, Semi-supervised, Inter-set correlation, Intra-set correlation

随着计算机技术、通信技术以及网络技术的飞速发展,在文本挖掘、生物信息技术、入侵检测等领域产生大量的高维(high dimension)数据。作为一种数据预处理技术,特征选择起着非常重要的作用:可以解决高维诅咒的问题,有效降维;通过删除无关或冗余特征改善机器学习效果;加速学习模型建立的过程。按照样本集中的实例是否有类标号,特征选择算法可分为有监督模型[1,2]和无监督模型[3]。很多学习算法(如文本分类)需要大量的标记样本(labeled sample),但已标记的样本能提供的信息有限;另一方面,容易获得的未标记样本(unlabeled sample)数量相对较多,且更接近整个样本空间上的数据分布。对样本进行标记往往需要缓慢的手工劳动,这制约了整个系统的构建,即所谓的标记瓶颈的问题。因此,如何通过少量的有类标记样本和大量的无类标记样本有效地进行特征选择,已成为目前关注的焦点[4]。

本文基于Relevant Set Correlation(RSC)模型[6],在半监督学习模式下提出了一种新颖的特征选择方法SFRSC(Semi-supervised Feature selection based on RSC model),从原始的、少量的已标记的训练样本出发,通过RSC模型寻找最相关的邻居点,将类标号自然地扩展到邻居点,并以最终形成的复合的聚类效果作为衡量特征子集的标准。

1 半监督特征选择的相关工作

Handl等在文献[12]中从多目标优化的角度进行半监督特征选择,并在实验中证明了在缺乏先验知识时,基于pareto的优化方法可以达到较好的效果。文献[7]基于图形理论,在半监督机器学习的同时考虑了特征选择的问题,首先选出所有两两相关的特征对,接着构造特征集的邻近图,通过寻找最大连通子图得到互相相关的特征子集。文献[9]中,作者提出的框架也是将类标号从少量的训练样本扩展到未标记实例上。主要思想是首先在原始训练集上得到初始分类器和特征子集,抽样选取一定数量的未标记实例,通过现有的分类器得到类标号后形成新的训练集。重复该过程直到特征子集达到一定的规模或迭代达到一定的次数。虽然文献[9]证明了该方法的有效性,但事实上并没有充分地利用未标记点的信息。文献[10]利用spectral图理论,构造特征向量,通过衡量该向量与数据(包括已标记和未标记)的符合程度对特征进行排序。算法是filter类型的,时间可能较长,并且与具体应用中涉及的学习方法无关。

2 RSC模型的基本概念

RSC模型最早是由Houle提出的[11],接着在文献[5]中给出了改进后的聚类算法。在此基础上Houle等人在文献[6]研究了如何无监督地选择特征,并实现了算法RSCF。现在的工作与之类似,但也有很大的不同:(1)RSCF仅仅利用了未标记实例,没有考虑先验知识。(2)RSCF是一种过滤模式算法,与具体的机器学习方法无关;而本文提出的算法是封

到稿日期:2008-12-03 返修日期:2009-03-02 本文受863国家重点基金项目(2006AA01Z451,2007AA01Z474,2007AA010502)资助。
王 博(1981—),女,博士生,主要研究方向为数据挖掘、网络安全等,E-mail:bowang_s.x@163.com;贾 焰(1960—),女,教授,博士生导师,主要研究方向为数据库和网络安全;田 李(1980—),男,博士,主要研究方向为数据流挖掘和网络安全。

装模式，以机器学习的效果作为衡量特征的标准。(3)RSCF的衡量方法需要考虑样本集中每一个实例的贡献。下面简单给出RSC模型的基本思想，以及在本文中用到的符号，如表1所列。

表1 符号表

S	某一领域的样本集
q	项
$\pi(q)$	对于项q，给出S中其它项的一个排序，当$i<j$时表示q_i与q更相似或相关
$\pi(q,i)$	表示排序结果的第i个项
$Q(q,k)$	表示排序结果的前k个项，即$Q(q,k)=\{\pi(q,i)\mid 1\leqslant i\leqslant k\}$

值得说明的是，当确定了特征子空间和相关度标准后，排序函数π也就确定了。当选择不同的特征子集，$\pi(q)$可能得到不同的结果，反映了两个样本间不同的相关程度。所以特征子集与样本集分布的符合程度，可以由$\pi(q)$间接地反映。下面给出文献[6]中样本集之间的相关度以及样本集自相关度的定义。

定义1 假设$A,B\subseteq S$，它们之间的相关度(用皮尔森系数衡量)可简化地表示为式(1)，其中$|\cdot|$表示集合中元素的个数：

$$R(A,B)=\frac{|S|\left(\frac{|A\cap B|}{\sqrt{|A||B|}}-\frac{\sqrt{|A||B|}}{|S|}\right)}{\sqrt{(|S|-|A|)(|S|-|B|)}} \tag{1}$$

集合A的相关度由A中所有元素决定。RSC基于这样的假设：对于任意$v\in A$，如果v与所有$v'\in A/v$都是强相关的，那么S中与v强相关的项都属于A。可以得到衡量集合A自相关度的(first-order)定义，如下：

定义2 $SR(A)\triangleq\frac{1}{|A|}\sum_{v\in A}R(A,Q(v,|A|))$，并且$SR(A)$为1表示$A$内部是强相关的，相反为0则表示$A$自身没有相关性。

可以看到，计算集合A的自相关度与排序函数π有密切的关系。文献[6]注意到定义2存在着不足：当π比较特殊时，$SR(A)$的计算结果是不太理想的。例如对于任意$v\in A$，π得到相同的排序结果，即$\pi(v)=\pi(v')$，这时得到的$SR(A)$没有任何意义。又例如当考虑样本集S时，不论π取怎样的值，$SR(S)$永远都等于1。为了解决上述这些问题，通过考虑A与$v\in S/v$的相关性，给出了更科学的计算方法[6]：

$$DR(A)\triangleq\frac{1}{|A|}\sum_{v\in A}R(A,Q(v,|A|))-\frac{1}{|S|-|A|}\sum_{v\notin A}R(A,Q(v,|A|)) \tag{2}$$

根据式(2)，若对于任意项π都返回相同的排序结果，那么$SR(A)$总是等于0。

3 SFRSC算法

基于上述工作，本节提出了一种新颖的半监督特征选择算法——SFRSC(Semi-supervised Feature selection based on RSC model)。首先对一些关键步骤进行说明，然后具体描述算法。

3.1 类标号的扩展

扩充训练集是半监督特征选择常用的方法之一[8]。本文也是基于这样的思想，通过RSC模型，以现有的已标记实例为核心点，将类标号扩展到未标记样本，这些被新标记的样本又可作为核心点继续向周围扩展。假设S中最初有t个已标记样本，记作p_i(其中$i=1,\cdots,t$)。若S中的样本可被标记为c类，那么$p_{i,l}(1\leqslant l\leqslant c)$表示第$i$个样本属于第$l$个类。首先通过式(3)得到$k$，那么包含$p_i$并且自相关度最大的集合就是$A=Q(p_i,k)$。也就是说，$A$中样本间的相关度最大，因此它们的类标号应该相同。由此可以确定A中未标记样本所属的类别，这些样本可进而作为新的核心点，继续将标号向外扩展。

$$\underset{1\leqslant k\leqslant S}{\operatorname{argmax}}\,DR(Q(p_i,k)) \tag{3}$$

图1—图4具体说明了整个扩展的过程。如图所示，空心圆点表示S中未标记的样本，实心点表示已标记样本，不同的形状代表不同的类。为了表述方便，类别用数字标明，例如实心三角形代表那些属于第2种类型的样本。图1表示初始情况下，样本集中包含4个已标记样本，分别属于3个不同的类别。以这4个样本为初始核心点，根据式(3)得到各自的A，A中未标记样本的所属类别与对应的核心点相同。这是第一步扩展，得到的结果如图2所示。图2得到一些新的带类标记样本，在图3中将其作为要考虑的核心点，进一步通过式(3)进行扩展，直到所有的样本都得到预测。该扩展过程存在这样的问题：最终的划分可能出现重叠，即某个(些)点被多个类标号标记。如图4中的p'(用星号表示)同时属于类1和类2，这是由于式(3)只能得到局部最优。本文假设一个样本仅能属于一个类型，所以图4中出现的情况是不允许的，将在4.2节中给出解决方案。

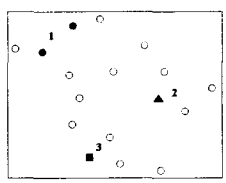

图1

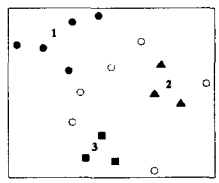

图2

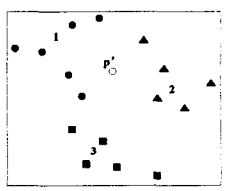

图3

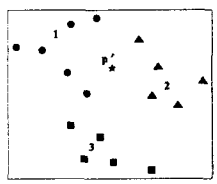

图4

3.2 对重叠部分的处理

假设样本v位于集合A和B(A,B中的样本分别属于不同类)的重叠区域，即$v\in A\cap B$。直观地，因为v是根据式(3)加入到集合A和B的，可计算v对集合自相关度的贡献来判断其归属。式(4)反映了单个样本对集合自相关度的影响，值越大表明贡献越大。分别计算$IR_1(v|A)$和$IR_1(v|B)$，若对A的贡献大，那么v的类型与A中样本的类型一致。反之v的类型与B中样本的一致。如果两个值相等，那么随机选择类型。

$$IR_1(v|A)=R(A,Q(v,|A|)) \tag{4}$$

式(4)仅考虑了样本v对单个集合自相关度的贡献，一旦确定了v的所属类别后，就没有考虑由于集合A,B的变化带来两者关系的变化。式(5)基于这样的考虑：将v归到集合A中，当且仅当v对A的自相关度贡献较大，且A和新产生的

集合 B'(B 去掉 v)的离散度增大(即就是 A 与 B' 的相关度减小)。根据式(5)分别计算 $IR_2(v|A,B)$和 $IR_2(v|B,A)$并进行比较。

$$IR_2(v|A,B)=R(A,Q(v,|A|))-R(A,B/v) \quad (5)$$

第 4 节中通过实验对两种不同的判断方法进行比较。如未作特别声明,本文的算法使用式(5)。为了方便描述实验,若算法使用式(4)则记作 SFRSC-IR1。

3.3 评价方法

本文选用封装类型的特征选择方法,通过最终样本集的划分结果评价特征子集。采用复合评价方法,同时考虑类自身的相似程度以及类间的离散程度。

对类标号进行扩展并对重叠部分的样本进行判断后,样本集 S 已被划分成互不相交的 m 个部分 $P=\{P_1,P_2,\cdots P_m\}$。根据上述分析,由式(6)对排序函数 π 进行评价。当确定了两个样本间相关性的衡量方法后,函数 π 的优劣就反应了相应特征子集的好坏。如果 $\phi(\pi_1)>\phi(\pi_2)$,π_i 对应的特征子集为 f_i,那么 f_1 要优于 f_2。

$$\phi(\pi)=\frac{1}{m}\sum_{1\leqslant i\leqslant m}SR(P_i)-\frac{1}{2m}\sum_{1\leqslant i,j\leqslant m,i\neq j}R(P_i,P_j) \quad (6)$$

3.4 算法描述

考虑到 S 中可能存在某些噪声点,当未标记样本的个数减少到某一数量时,循环结束。首先给出算法 SFRSC-P(SFRSC-Partititon)的描述,当特征子集确定后,SFRSC-P 输出由该子集得到的 S 的划分结果,如表 2 所列。在表 3 中,SFRSC 通过 SFFS(sequential forward feature selection)选择具体的特征子集,将 SFRSC-P 作为子过程得到相应划分结果并衡量该子集,迭代若干次后最终得到满足用户需求的特征子集。

表 2 算法 SFRSC-P

算法 1 SFRSC-P

输入:ini_L,ini_UL
输出:a partition of S, $P=\{P_1,\cdots,P_m\}$
extendedL:新生成的已标记样本集
toExtendL:*extendedL* 中原来未标记的样本集
class(v):v 属于的类别

currentL:目前的已标记样本集,*currentUL*:目前的未标记样本集

$P_1\cdots P_m\leftarrow\emptyset$, *extendedL*$\leftarrow\emptyset$, *newBornL*$\leftarrow\emptyset$;
currentL$\leftarrow$*ini_L*, *currentUL*$\leftarrow$*ini_UL*;
toExtendL$\leftarrow$*ini_L*;
repeat
for every $p\in$ *toExtendL* do
  $k=\underset{1<k<S}{\operatorname{argmax}}DR(Q(p,k))$;
extendedL$\leftarrow Q(p,k)$;
$L'\leftarrow$*extendedL*$-$*currentL*;
$UL'\leftarrow$*extendedL*$-$*currentUL*;
for every $v\in UL'$ do
  class(v)$\leftarrow$*class*(p);
  j=*class*(v);
  $P_j\leftarrow P_j+\{v\}$;
end
for every $v\in L'$ do
  if *class*(v)$\neq$*class*(p) then
    if $IR_1(v|class(v))<IR_1(v|class(p))$(or $IR_2(v|class(v),class(p))<IR_2(v|class(p),class(v))$)
  then *class*(v)$\leftarrow$*class*(p);
    t=*class*(p);
    j=*class*(v);
    $P_j=P_j-\{v\}$;
    $P_t=P_t+\{v\}$;
  end
newBornL$\leftarrow$*newBornL*$+UL'$;
currentL$\leftarrow$*currentL*$+UL'$;
currentUL$\leftarrow$*currentUL*$-UL'$;
end
toExtendL$\leftarrow$*newBornL*;
newBornL$\leftarrow\emptyset$;
until $|currentUL|\leqslant\varepsilon$
return $P=\{P_1\cdots P_m\}$

表 3 算法 SFRSC

算法 2 SFRSC

输入:*ini_fs*, *ini_L*, *ini_UL*　　输出:*best_fs*
currentFS:目前评价的特征子集
best_fs:最终选择的特征子集
for i=1 to *ini_fs* do
Use algorithm SFFS to decide the i-th feature in currentFS on *ini_fs* with $\phi(\pi)=\frac{1}{m}\sum_{1\leqslant i\leqslant m}SR(P_i)-\frac{1}{2m}\sum_{1\leqslant i,j\leqslant m,i\neq j}R(P_i,P_j)$ as the evaluation criteria, where $P=\{P_1\cdots P_m\}$=SFRSC-P(*ini_L*$\circ$*currentFS*, *ini_UL*$\circ$*currentFS*);
if $\phi(\pi)\geqslant\lambda$ or *Size*(*currentFS*)=SIZE
then break;
end
best_fs$\leftarrow$*currentFS*;
return best_fs.

"$A\circ FS$"指的是集合 A 在特征集 FS 上的投影。ε 表示当前 S 中仍未标记样本的个数,当 ε 达到一定的数目时 SFRSC-P 中的循环结束。当 $\phi(\pi)\geqslant\lambda$(表明对应的特征子集已得到较好的划分结果),或当 *Size*(*currentFS*)等于用户给定的 SIZE 时,SFRSC 的循环停止。

4 实验结果

为了验证算法 SFRSC 的有效性,本文在若干 UCI[1] 的数据集上进行实验,包括 Ionoshpere, glass, sonar, ecoli 以及 wine。同时还对 KDDCUP 99 入侵检测数据[2] 进行抽样得到 20995 个实例(其中 19.69%的实例是正常实例),该子集中类的分布情况与原样本集中大致相同。每个样本集都被分为 3 部分,10%作为已标记样本,45%作为未标记实例,剩余的用作测试集。为了比较,本文还选取了算法 ReliefF(有监督特征选择算法)和 Laplacian Score(无监督特征选择算法,用 LS 代表)。用 1-NN 分类器的精度来衡量算法的有效性。

所有的实验都基于一个操作系统为 windows XP,内存为 512M,CPU 为 2.0 GHz 的 pc 机。算法 SFRSC 中,参数的值设为 ε=1%*(未标记样本的个数),$\lambda=\infty$,SIZE 值在各具体

(下转第 208 页)

[1] http://www.ics.uci.edu/?mlearn/MLRepository.html

[2] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

rence on Artificial Intelligence. 1999:318-325
[9] Kautz H, McAllester D, Selman B. Encoding Plans in Propositional Logic[C]// Proceedings of the 4th International Conference on Knowledge Representation and Reasoning. 374-385
[10] Blum A L, Furst M L. Fast planning through planning graph analysis[J]. Artificial Intelligence, 1997, 90:281-300
[11] 姜云飞,吴康恒. 智能规划的研究和应用[J]. 计算机科学,2002,2(29):100-103

(上接第 191 页)

实验中被给出。表 4 给出各数据集的描述以及各算法的实验结果,其中标黑的数据表示最优结果。其中,"Full"表示不考虑特征选择的算法。

由表 4 可以看出,大多数据集中 SFRSC 的分类精度都高于其它算法。特别地,在所有数据集中 SFRSC 的效果都要比 LS 好,证明了先验知识在特征选择中的作用。而 Full 算法却在 glass 和 ecoli 数据集中的效果更为显著,可能是由于特征个数较少的原因。

表 4 各个数据集信息及各个算法的结果

Dataset	Instance	Dimesion	Class	Full	ReliefF	LS	SFRSC
Iono	351	34	2	73.17	74.53	67.82	**76.31**
glass	214	10	7	**65.44**	62.73	58.12	60.91
sonar	208	60	2	58.86	67.21	67.73	**71.04**
ecoli	336	8	8	**70.52**	**70.47**	67.16	66.58
wine	178	13	3	77.02	78.35	62.05	**89.15**
KDD	20995	42	5	62.49	71.04	71.39	**73.55**

为了进一步与其它方法进行比较,实验还考虑了随着已标记样本个数或特征子集大小的变化,基于各算法得到的特征子集,分类精度的变化情况。图 5 以数据集 wine,sonar 为代表说明特征子集的大小对分类精度的影响,可以看出 SFRSC 基本上都是优于 ReliefF 和 LS 的。由于选取的已标记样本较少,该对比结果体现了 SFRSC 能充分利用标记样本和未标记实例携带的信息。随着所选特征数目的增加,各算法的性能有所提高。当接近于原始特征集大小时,算法的性能会有所降低。该现象在 wine 数据集中更明显,这可能是由于后期加入的 feature 对精度的影响。

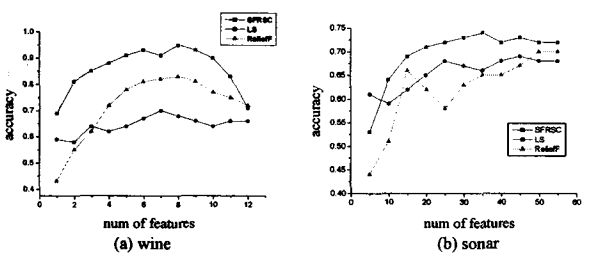

图 5 特征子集的大小对分类精度的影响

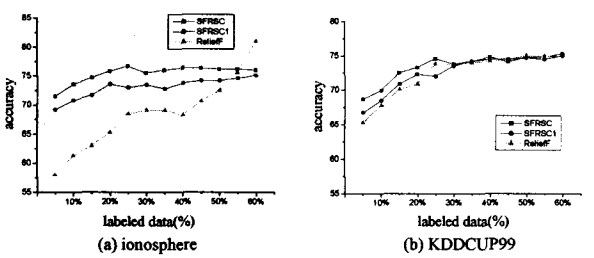

图 6 已标记样本集合的大小对分类精度的影响

图 6 在数据集 Ionosphere 和 KDDCUP99 上给出不同大小的标记样本集与分类精度的关系。特征数目固定取 11。由于算法 LS 的性能与标记样本的个数无关,这里考虑的是 SFRSC-1(与 SFRSC 的区别仅在于对重叠部分的处理不同)。如图 6 所示,总体来说增加标记样本的个数可以提高算法的精度。当标记样本较少时,精度的增长较快,并且 SFRSC 和 SFRSC-1 的优势也较为明显。但标记样本占整个数据集的比例超过一定阈值(大约是 25%)时,曲线变得比较平缓,精度增长减慢。所以过多的先验知识并不能过快地提高分类精度,SFRSC 可以基于小标记样本集得到较好的学习效果。另外,对于相同大小的标记样本集,SFRSC 总是优于 SFRSC-1,这也证明了式(5)优于式(4)。

结束语 在已标记样本有限的前提下,本文基于 RSC 模型提出了一种半监督特征选择算法 SFRSC。该算法是一个迭代的过程,在特征子集对应的空间中,将已标记样本作为核心点,将类标号扩展到未标记样本上,并以 S 的最终划分的复合结果作为衡量该特征子集的标准。同时还考虑了一个样本被标记为多个类的情况。实验证明,该算法是比较有效的。

本文的工作基于这样的假设:初始的已标记样本集中包含所有的类别信息。下一步将考虑在确定类标号的样本集中,某些类别没有对应的样本。另外,本文考虑的先验知识是类别信息,但往往样本之间的约束关系(must-link 和 cannot-link)更容易得到,在这种情况下考虑特征选择也是非常有意义的。

参 考 文 献

[1] Yang K, Yoon H, Shahabi C. A Supervised Feature Subset Selection Technique for Multivariate Time Series
[2] Liu H, Yu L. Toward Integrating Feature Selection Algorithms for Classification and Clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(4):491-502
[3] Zhao Zheng, Liu Huan. Searching for Interacting Features[C]// ijcai 2007
[4] Seeger M. Learning with labeled and unlabeled data[R]. 2000
[5] Houle M E. Clustering without data : the GreedyRSC heuristic [C]//Proc. International Workshop on Data-Mining and Statistical Science(DMSS 2006). Sapporo, Japan, September 2006:62-69
[6] Houle M E, Grira N. A Correlation-Based Model for Unsupervised Feature Selection[C]//CIKM'07
[7] Izutani A, Uehara K. A Modeling Approach Using Multiple Graphs for Semi-Supervised Learning[J]. Discovery Science, 2008:296-307
[8] Nakatani Y, Zhu K, Uehara K. Semisupervised learning using feature selection based on maximum density subgraphs[J]. Systems and Computers in Japan(SCJAPAN), 2007, 38(9):32-43
[9] Ren Jiangtao, Qiu Zhengyuan, Fan Wei, et al. Forward Semi-Supervised Feature Selection[C]//PAKDD08. 2008
[10] Zhao Z, Liu H. Semi-supervised Feature Selection via Spectral Analysis[C]//SIAM International Conference on Data Mining (SDM-07). 2007
[11] Houle M E. Clustering without data: the relevant set correlation model[C]//Proc. International Workshop on Data-Mining and Statistical Science (DMSS 2006). Sapporo, Japan, September 2006:54-61
[12] Handl J, Knowles J. Semi-supervised feature selection via multiobjective optimization[C]//IJCNN06. 2006