一种改进离散度的特征选择方法①

兰远东[1,2]，邓辉舫[2]

[1](惠州学院 计算机科学系，惠州 516007)
[2](华南理工大学 计算机科学与工程学院，广州 510006)

摘 要：降维在机器学习中起着至关重要的作用。而降维的方法主要有两类：特征选择和特征提取。离散度方法是特征选择中常用的一种方法，通过计算每个特征的离散度来选择特征，被选择的特征一般都具有较高的离散度值。但是离散度的计算没有考虑到特征间的相互影响。通过改进离散度的计算，不单考虑到类间相同特征对离散度的影响，还考虑到不同特征之间的离散度影响。在 UCI 数据集上的实验证明，改进离散度的特征选择具有较好的性能。

关键词：特征选择；机器学习；离散度；模式分类；特征提取

Feature Selection Method Based on Improved Scatter Degree

LAN Yuan-Dong[1,2], DENG Hui-Fang[2]

[1](Department of Computer Scienne, Huizhou University, Huizhou 516007, China)
[2](School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China)

Abstract：Dimension reduction is important in machine learning. The two methods of dimension reduction are feature extraction and feature selection. Scatter degree is one of the feature selection methods which attribute a degree of scattering for each feature. Features are selected that have higher scatter degree. In this paper, classification error has been reduced by considering other aspects in computing scatter degree. Experiments on UCI dataset show that improved scatter degree have a good performance on feature selection.

Key words：feature selection; machine learning; scatter degree; pattern classification; featre extraction

降维（Dimension reduction）是指将高维数据按照某种方法以最小信息损失投影到一个低维子空间中，是机器学习、数据挖掘和模式识别领域常用的方法，其目的是为了使得相应的算法运行更快。降维实际上就是特征约减，特征约减主要有以下两种方法：特征选择和特征提取[1]。

现在已存在多种特征提取的算法，它们大多基于原始特征的组合或变换来得到低维的新特征。比较常用的一些算法有主成分分析（Principal Component Analysis, PCA）[2]，线性判别分析（Linear Discriminant Analysis, LDA）[3],潜在语义分析（Latent Semantic Analysis, LSA）[4]，局部线性嵌入（Locally Linear Embedded, LLE）[5]等。

特征选择是指从原始特征中选择（不使用特征组合和特征变换）一个特征子集，常用的算法有基于支持向量机的方法（Support Vector Machine, SVM）[6]，信息增益和 χ^2-测试[7]，以及基于离散度（Scatter degree, SD）的方法[8]等。

相对与特征选择来说，特征提取通过特征组合和特征变换的方法获得的新特征通常具有更好的判别特性，但是通过特征组合和特征变换得来的新特征却可能不具有实际的物理意义。而特征选择算法通常具有简单易行，计算速度快，测量代价低和能保持原始物理意义等特点。

基于离散度的特征选择中，每一个特征的重要程度（或分类权重）主要取决于“离散度”。在本文中，

① 基金项目:国家自然科学基金(61170193)
收稿时间:2011-10-21;收到修改稿时间:2011-11-20

我们给出了一种改进离散度的特征选择方法（A Feature Selection Method based on Improved Scatter Degree），在计算每个特征的离散度时，同时考虑到相同特征在类间的均值距离和不同特征之间的相互影响。

1　离散度特征选择

使用离散度的特征选择方法主要基于线性判别分析（LDA）使用 Fisher 判别准则，将样本间的离散度归结于样本的每一个特征，通过下面的公式计算每一个特征的离散度：

$$SD(t_i)=\frac{S_B(i,i)}{S_W(i,i)} \tag{1}$$

在式（1）中是 t_i 第 i 个特征；S_W 是类内的散度矩阵，其计算如式（3）所示；S_B 是类间散度矩阵，其计算如式（2）所示。

$$S_B=\sum_{i=1}^{c}n_i(m_i-M)(m_i-M)^T \tag{2}$$

$$S_W=\sum_{i=1}^{c}\sum_{j=1}^{n_i}(d_{ij}-m_i)(d_{ij}-m_i)^T \tag{3}$$

$$m_i=\frac{1}{n_i}\sum_{j=1}^{n_i}d_{ij} \tag{4}$$

$$M=\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n_i}d_{ij} \tag{5}$$

在式(2)～(5)中，M 是所有特征的均值向量；m_i 是第 i 类的特征均值向量；d_{ij} 是第 i 类的第 j 个样本的向量表达；n_i 是第 i 类的样本数量。

在计算完每一个特征的离散度（SD）之后，将每一个特征按照离散度从大到小排列，特征选择的过程就是从离散度较大的特征开始，从大到小选出一个合适特征子集。也就是说，我们在对样本做特征选择的时候，是根据样本的每一个特征的离散度的大小来决定该特征的重要程度。

2　改进离散度的特征选择

2.1 改进离散度

在式(1)中，计算每一个特征的离散度的时候，只是对每一个特征独立计算其离散度，然后做特征选择的时候，根据计算得到的每一个特征的离散度值来决定每一个特征的重要程度。也就是在计算离散度时只关注某个类中第 i 个特征的均值与其他类中相应特征均值之间的距离（类间相同特征均值距离），以此来作为分类的标准，完全没有考虑到其他特征对该特征离散度的影响。

然而，在大多数分类问题中，类间相同特征均值的距离和不同特征之间的距离同样重要。因此，我们想在基本的离散度特征选择方法中融入其他特征对离散度的影响，以改进离散度的计算方法。

$$ISD(t_k)=\frac{\sum_{i=1}^{c}n_i\sum_{j=1}^{l}[(m_{ik}-m_{ij})-(M_k-M_j)]^2+S_B(k,k)}{S_W(k,k)} \tag{6}$$

其中 $\sum_{i=1}^{c}n_i\sum_{j=1}^{l}[(m_{ik}-m_{ij})-(M_k-M_j)]^2$ 是加入的类间特征之间的距离影响；m_{ik} 是第 i 类中第 k 个特征的均值向量；M_k 是所有样本的第 k 个特征的均值向量。

基于式（6），ISD 越大的特征，其对分类的影响就越大，也就是特征选择是优先选择的特征。

2.2 基于改进离散度的特征选择算法

在改进离散度后的特征选择时，我们先计算每一个特征的 ISD，然后根据 ISD 的大小排序，其值大的相应的重要性也就越大。因此，当我们知道需要将特征空间投影到多少维度的子特征空间时（比如 L 维子空间），就根据 ISD，采用从大到小的原则选择前面 L 个特征，其具体的特征选择算法如下：

算法名称：改进离散度的特征选择

输入：数据集，特征子空间的维度 L

输出：数据集的子特征表示

1. 根据式（4）和（5）计算 M 和 mi
2. 根据式（2）和（3）计算 SB 和 SW
3. For i=1 to n　　//n 为样本维度
4. 根据式（6）计算每一个特征的 ISD(ti)
5. End for
6. 根据 ISD(ti)（从大到小）对特征排序
7. 选择前 L 个特征，作为数据集的子特征表示

从上面的伪代码，我们可以看出：对于数值型特征，该算法非常容易实现，对于非数值型特征，只要具有相应的特征比较方法，算法也是比较容易实现。

3　实验结果与分析

在实验中我们采用 KNN(K Nearest Neighbors)方

法来实现我们提出的特征选择方法，其中 K 最后选定为 5。

算法名称：改进离散度的特征选择

输入：数据集，特征子空间的维度 L

输出：数据集的子特征表示

1. 根据式（4）和（5）计算 M 和 m_i
2. 根据式（2）和（3）计算 S_B 和 S_W
3. For i=1 to n //n 为样本维度
4. 根据式（6）计算每一个特征的 $ISD(t_i)$
5. End for
6. 根据 $ISD(t_i)$（从大到小）对特征排序
7. 选择前 L 个特征，作为数据集的子特征表示

3.1 实验数据集

在实验中我们选择国际通用的 UCI（University of California Irvine）[9]数据集，这些数据集可以直接从 UCI 的网站上下载。我们选择的 10 个数据集如表 1 所示：

表 1 用于实验的 10 个 UCI 数据集

Datasets	The number of	
	documents	features
Arcene	900	10000
20newsgroups	3960	8014
HIVA	3835	1617
GINA	3153	970
ISOLET	7797	617
SECOM	1567	591
Madelon	4400	500
Semeion Handwritten Digit	1593	256
SYLVA	13076	108
ADA	41461	48

3.2 评价标准

我们使用 F 值度量（查准率 precision(P)和查全率 recall(R)）来评价特征选择的性能，P 和 R 的计算，如式（7）和式（8）所示：

$$P_i = \frac{TP_i}{TP_i + FP_i} \tag{7}$$

$$R_i = \frac{TP_i}{TP_i + FN_i} \tag{8}$$

其中，P_i 是第 i 类的查准率；R_i 是第 i 类的的查全率；TP_i 是第 i 类中正确选择的特征数；FP_i 是第 i 类中错误选择的特征数；FN_i 是第 i 类样本的错误分类数。

我们计算两个平均值，分别是微平均（Micro Average）和宏平均（Macro Average）。对宏平均的计算，必须对每一个类计算相应的查全率和查准率，然后加起来除以总的类数；微平均通过式（9）和式（10）计算。

$$F_i = \frac{2P_iR_i}{P_i + R_i} \tag{9}$$

$$F(Macro-average) = \frac{1}{c}\sum_{i=1}^{c} F_i \tag{10}$$

3.3 实验结果

如上所述，我们分别在 10 个数据集上做实验验证改进离散度的特征选择的有效性，为了表示简洁，只给出 3 个数据集上（Arcene, 20newsgroups, GINA）的实验结果，其他数据集上也有相似的结论。我们将改进离散度的特征选择方法与基本的离散度特征选择方法最对比实验，在特征选择的时候，子特征空间在原始特征空间的 10%到 90%之间变换，在 3 个数据集上的实验结果如图 1-3 所示（分为微平均（Micro Average）和宏平均（Macro Average））。

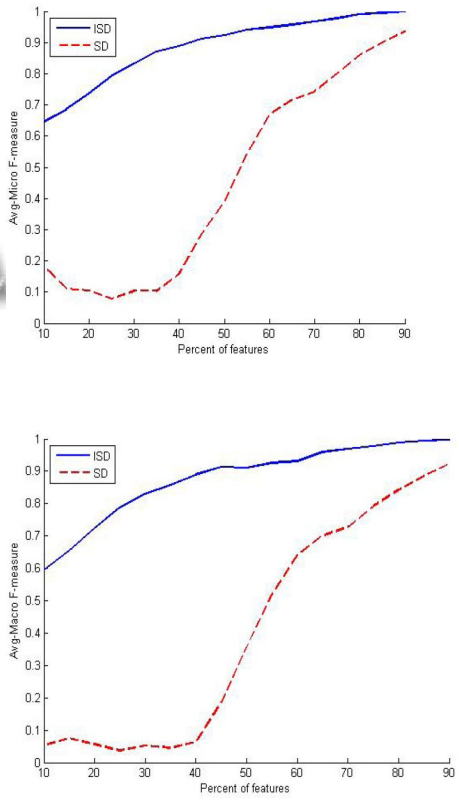

图 1 在 Arcene 数据集上的性能对比

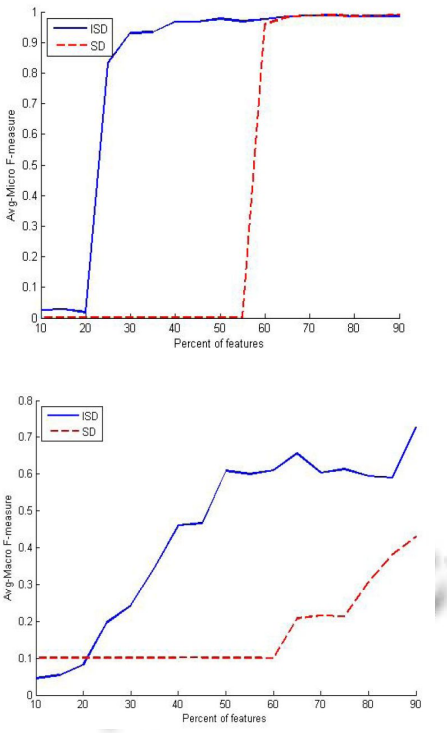

图 2　在 20newsgroup 数据集上的性能对比

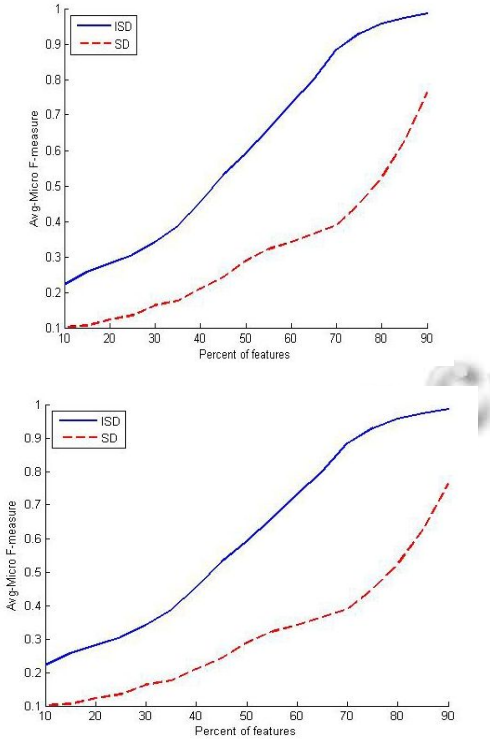

图 3　在 GINA 数据集上的性能对比

从图 1-3 所示的结果可以看出，改进离散度的特征选择方法比基本的离散度特征选择方法具有更好的性能。

4　结论

在本文中，我们给出了一种改进离散度的特征选择方法，由于在计算特征离散度的时候考虑到了特征间的相互影响，特征选择更为合理和准确。在 UCI 数据集上的实验表面，改进离散度的特征选择具有较好的性能，而且算法非常容易实现。

参考文献

1 周丽丽,李凡长.基于范畴的数据降维方法.计算机科学, 2011,9:242−245.

2 Martinez AM, Kak AC. PCA versus LDA. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2011,23(2):228−233.

3 Ye JP, Janardan R, Park CH, Park H. An optimization criterion for generalized discriminant analysis on undersampled problems. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2004,26(8):982−994.

4 Deerwester S, Dumais S, Furnas G, Landauer T, Harshman R. Indexing by latent semantic indexing. Journal of the American Society for Information Science, 1990,41(6): 391−407.

5 Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science. 2000,290(22):2323−2326.

6 Kim H, Howland P, Park H. Dimension reduction in text classification with support vector machines. Journal of Machine Learning Research, 2005,6(1):37−53.

7 Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. Proc. of the 15th International Conference on Machine Learning. Nashville, Tennessee, 1997. 412−420.

8 Xu JL, Xu BW, Wang C, Cui ZF. Feature selection based on scatter degree. Proc. of the International Conference on Machine Learning. Las Vegas, Nevada, 2008. 417−422.

9 http://archive.ics.uc.