计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(4):157−162 [doi: 10.15888/j.cnki.csa.006859]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

改进 PM-MD 的分类器集成①

吴立凡, 何振峰

(福州大学 数学与计算机科学学院, 福州 350116)
通讯作者: 吴立凡, E-mail: 1099123648@qq.com

摘　要: Hub 会对高维数据分析产生显著消极影响, 现有研究分别采用了五种降 Hubness 策略以提高分类效果, 但单个降 Hubness 策略适用范围有限. 为解决这一问题, 提出对五种降 Hub 分类器进行基于 PM-MD 的集成, PM-MD 集成是通过 KNN 确定分类对象的决策表以及通过分类器得到分类对象的类支持向量, 最后通过比较决策表和类支持向量的相似性计算分类器的竞争力权重. 由于 PM-MD 在处理高维数据集时高斯势函数存在弱化距离导致区分度不足的倾向, 因此提出了采用欧氏距离直接计算决策表以提高分类精度. 在 12 个 UCI 数据集上的实验结果表明: PM-MD 不仅获得更好且稳定的分类结果, 而改进后的 PM-MD 则进一步提高了分类精度.

关键词: PM-MD; 分类器; 集成; Hubness; 高维

引用格式: 吴立凡,何振峰.改进 PM-MD 的分类器集成.计算机系统应用,2019,28(4):157–162. http://www.c-s-a.org.cn/1003-3254/6859.html

Improved PM-MD Classifier Integration

WU Li-Fan, HE Zhen-Feng

(School of Mathematics and Computing Science, Fuzhou University, Fuzhou Fujian 350116, China)

Abstract: Hubs have a significant negative impact on high-dimensional data analysis. The current research uses five kinds of Hubness strategies to improve the classification effect, but each strategy has a limited scope of application. In order to solve this problem, PM-MD-based integration is proposed for the Hubness-based classifiers. The PM-MD integration determines a decision profile of the classification object through KNN and determines a class support vector of the classification object through the classifier. Finally, the competitiveness of the classifier integration is evaluated by comparing the similarity between the decision profile and the class support vector. When PM-MD dealing with high-dimensional data sets, because the Gaussian potential function tends to reduce the distance which leads to a lack of discrimination, it is proposed to use Euclidean distance to directly calculate the decision profile to improve the classification accuracy. The experimental results on 12 UCI datasets show that PM-MD obtains sound and stable classification results, and the improved PM-MD further improves the classification accuracy.

Key words: PM-MD; classifier; integration; Hubness; high dimensions

Hub[1]指一些经常出现在其他数据点的最相邻列表中的数据点, 它是随着维度的增加而出现的, 这种现象一般称为 Hubness 现象[2]. Hubness 是高维空间数据分布的一个固有性质, 对高维数据分析产生了显著的影响[2],受影响的方法和任务包括贝叶斯建模, 近邻预测和搜索, 神经网络等等[2]. 比如, Hubness 的出现会直接影响到 KNN 分类的准确性[3], 这是因为: Hub 在相应的距离空间中传播其编码信息过于广泛, 而非

① 基金项目: 福建省自然科学基金 (2018J01794)
Foundation item: Natural Science Foundation of Fujian Province (2018J01794)
收稿时间: 2018-10-23; 修改时间: 2018-11-14; 采用时间: 2018-11-22; csa 在线出版时间: 2019-03-28

Hub 携带的信息基本上丢失,导致这些距离空间不能很好地反映类信息[4].

为了减少 Hubness 的负面影响, 有两类 (共五种) 降 Hubness 的分类器策略应用于数据转换以提高 KNN 分类精度: 其中第一类策略 (二次距离策略) 将距离矩阵换算到二次距离 (NICDM[1]、MP[1]), 第二类策略 (线性换算策略) 直接应用于数据向量 (CENT[3]、MINMAX[5]、ZSCORE[5]). 第一类策略中: NICDM 是将 Hubness 问题与使最近邻关系对称的方法联系起来, 它需要得到每个数据点的局部邻域, 因此并不适用于非常大的数据库; MP 是一种无监督的将任意的距离函数转换成概率相似性 (距离) 测量的方法, 适用于非常大的数据库, 并且支持多个距离空间的简单组合. 实验表明 NICDM 和 MP 显著的减少了 Hubness, 提高了 KNN 分类精度, 还增强了近邻的稳定性[1,6]; 但是, NICDM 和 MP 适用于中心性和内在维度较高的数据集, 否则性能不稳定[1]. 第二类策略中: CENT 是将特征空间的原点移到数据中心, 可用于内积相似空间以减少 Hubness(其中每个样本和中心的相似性在 CENT 后等于零), 但它并不是适用于所有数据集, 它成功的必要条件是 Hub 与数据中心点有着高相似度[1]; 而 MINMAX 和 ZSCORE 则是应用很广泛的数据标准化方法, MINMAX 是对原始数据进行线性变换, 适用于原始数据的取值范围已经确定的情况; ZSCORE 基于原始数据的均值和标准差进行数据的标准化, 适用于最大值和最小值未知的情况, 或有超出取值范围的离群数据的情况.

在本文中对这两类策略进行多分类器集成, 由于不同的分类器提供了关于被分类的对象的补充信息, 因此多分类器系统可以获得比整体中任何单一的分类器更好的分类精度. 本文中的集成采用了一种计算特征空间分类器竞争力的名为 PM-MD[7]的方法. 在该方法中, 首先使用 KNN 的验证对象来确定分类对象 x 点的决策表, 决策表提供了被识别对象属于指定类的机会. 在概率模型中, 决策表的自然概念是基于 x 点上的类的后验概率. 接下来,将决策表与分类器所产生的响应 (支持向量或判别函数的值) 进行比较, 并根据相似性规则[8]计算分类器的分类竞争力:对决策表的响应越接近, 分类器的竞争力就越强[9,10].

本文的第 1 部分介绍两类降 Hubness 策略 (共五种), 第 2 部分介绍 PM-MD 集成的方法并对第 1 部分的策略进行集成, 第 3 部分介绍对 PM-MD 集成进行改进的部分, 第 4 部分对实验结果进行分析.

(1) 两类降 Hubness 策略

1) 二次距离策略

① NICDM

NICDM (Non-Iterative Contextual Dissimilarity Measure) 用 K 近邻的平均距离来重新换算距离, 相比于利用 K 近邻距离来重新换算距离, 这将返回更稳定的换算数据. NICDM 通过式 (1) 得到二次距离:

$$NICDM(d_{x,y}) = \frac{d_{x,y}}{\sqrt{\mu_x \mu_y}} \tag{1}$$

其中, μ_x表示 x 最近邻的平均距离, μ_y同理. NICDM 倾向于通过换算的数据点 x 和 y 的局部距离统计数据使得近邻关系更加对称[6].

② MP

相互接近 (Mutual Proximity) 通过将两个对象之间的距离转换为一个相互接近的距离来重新解释原始的距离空间, 使得两个共享最近邻的对象之间的距离就更紧密了, 而不共享最近邻的对象则相互排斥. 在 n 个对象集合中, 计算一个给定距离$d_{x,y}$可以归结为简单地计算出 j 与 x 和 y 之间大于$d_{x,y}$的距离[1]:

$$MP(d_{x,y}) = \frac{|\{j : d_{x,j} > d_{x,y}\} \cap \{j : d_{y,j} > d_{y,x}\}|}{n} \tag{2}$$

式 (2) 中, MP 是计算点 x 和 y 的相似度, 通过计算 1–MP 将相似度变成了二次距离[6].

2) 线性换算策略

① CENT

定心 (Centering) 将特征空间的原点转移到数据中心$c = (\frac{1}{|D|}) \sum_{x \in D} x$, 它是一种去除数据中观测偏差的经典方法, 但直到最近才被确定为减少 Hubness 的方法.

$$sim^{CENT}(x;q) \equiv < x - c, q - c > \tag{3}$$

式 (3) 中$sim^{CENT}(x;q)$是计算相似度, 需要通过计算1 – $sim^{CENT}(x;q)$将相似度变成了距离.

② MINMAX

在 MINMAX 算法里, 原始数据是线性变化的. MINMAX 使用式 (4) 将 v 值进行映射到v':

$$v' = \frac{v - x_{\min}}{x_{\max} - x_{\min}} \tag{4}$$

将$x_{\min}$和$x_{\max}$定义为样本中变量的最小值和最大值, MINMAX 将在[$x_{\min}$,$x_{\max}$]区间的训练样本映射到[0, 1].

③ ZSCORE

ZSCORE 通过式 (5) 将 v 值进行映射到v':

$$v' = \frac{v - \bar{x}}{\sigma_x} \tag{5}$$

其中, $\bar{x}$和σ_x分别为训练集中变量值的均值和标准差. 在映射之后, 平均值将为 0, 标准差将为 1.

(2) 集成方法

本文采用的 PM-MD 是一种全新的计算特征空间中分类器能力的集成方法, 被集成的方法为第一章中提到的 5 种策略: NICDM、MP、MINMAX、ZSCORE、CENT. PM-MD 方法是由两个方法结合起来: PM 方法 (Potential function Method) 和 MD 方法 (Max-max Distance). PM-MD 的特色在于对验证集的不同使用, 在 PM-MD 中验证集是用来评估测试点的类支持向量的, 并且在 PM 中使用 K 近邻来确定测试点的决策表, 最后在 MD 中由类支持向量与决策表的相似性决定分类器的竞争力[7].

集成流程的图示如图 1, 本文采用的是 5 种降 Hubness 策略和 KNN 分类, 也可以采用其他策略和分类方法.

1) 类支持向量

分类器ψ_l相当于一个从特征度量空间$x \subseteq R^{\dim}$到一个类标签集合$M = \{1,2,\cdots,m\}$的函数. 对于每个数据点 x, 分类器ψ_l经过该分类器的数据转换后通过 KNN 找到 x 的 K 近邻从而生成相应的类支持向量:

$$d_l(x) = [d_{l1}(x), d_{l2}(x), \cdots, d_{lm}(x)] \tag{6}$$

其中, $d_{lj}(x) \geqslant 0, \sum_{j \in M} d_{lj}(x) = 1$, $d_{lj}(x)$代表数据点 x 的 K 近邻中第 j 类的比重.

2) 根据 PM 计算决策表

决策表$\omega(x) = [\omega_1(x), \omega_2(x), \cdots, \omega_M(x)]$提供了数据点 x 属于指定类的机会, 其中$\omega_j(x) \geqslant 0, \sum_{j \in M} \omega_j(x) = 1$.

用 PM 方法通过 K 近邻计算数据点 x 的决策表的步骤如下:

① 计算一个非负势函数$H(x, x_k)$[11], 其值随着 x 和x_k之间距离增大而快速减少 (x_k为来自验证集的数据对象):

$$H(x, x_k) = \exp(-\|x - x_k\|) \tag{7}$$

② 根据上一步得到的势函数计算决策表$\omega_j(x)$, 它是 x 属于第 j 类的机会的衡量:

$$\omega_j(x) = \frac{\sum_{x_k \in N_K(x): j_k = j} \exp(-\|x - x_k\|)}{\sum_{j \in M} \sum_{x_k \in N_K(x): j_k = j} \exp(-\|x - x_k\|)} \tag{8}$$

其中, $N_K(x)$为验证集 V 中的数据点 x 的 K 近邻集合, x_k为来自验证集的数据对象, j_k为相应的类标签.

3) 根据 MD 计算分类器竞争力

分类器竞争力$c(\psi_l|x)$用来衡量分类器ψ_l在数据点 x 的分类能力, 它随着支持向量$d_l(x)$和决策表$\omega(x)$的距离$dist[\omega(x), d_l(x)]$的增大而减少.

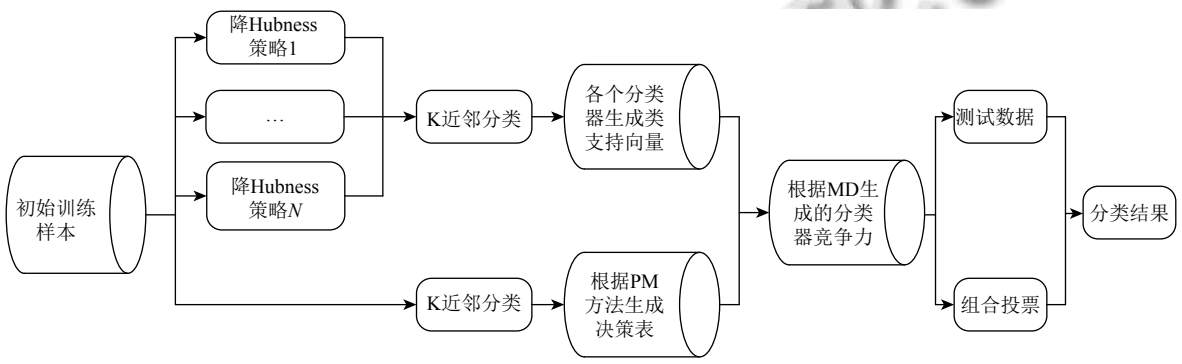

图 1 结合降 Hubness 过程的分类器集成框架

根据 MD 方法计算分类器竞争力步骤如下:

① 令 j 为分类器ψ_l在数据点 x 产生的类支持向量的最优值的类编号,即$d_{lj}(x) = \max_{k \in M}(d_{lk}(x))$; 同理, 令 i 为决策表$\omega(x)$在数据点 x 产生的最优值的类编号.

则支持向量$d_l(x)$和决策表$\omega(x)$的距离定义如下:

$$dist[\omega(x), d_l(x)] = |d_{lj}(x) - \omega_j(x)| + |d_{li}(x) - \omega_i(x)| \tag{9}$$

假设类支持向量为$d_l(x) = [0.1, 0.4, 0.2, 0.3]$, 决策表为$\omega(x) = [0.2, 0.1, 0.2, 0.5]$, 则$d_{lj}(x) = 0.4, j = 2$, $\omega_i(x) = 0.5, i = 4$. 所以距离计算如下:

$$dist[\omega(x), d_l(x)] = |0.4 - 0.1| + |0.3 - 0.5| = 0.5 \tag{10}$$

② 根据上一步得到的距离计算竞争力$c(\psi_l|x)$:

$$c(\psi_l|x)=1-\frac{|d_{lj}(x)-\omega_j(x)|+|d_{li}(x)-\omega_i(x)|}{2} \tag{11}$$

4) 组合投票以及最后分类精度的计算

对于测试数据点 x, 其最后的分类结果$\psi(x)$是由分类器组合中的每个分类器产生的类支持向量 (式 (6)) 结合其对应的分类器竞争力 (式 (11)) 组合投票得出来的:

$$D_j(x)=\sum_{l=1}^{T}c(\psi_l|x)d_{l,j}(x) \tag{12}$$

其中, T 为分类器的个数.

$$\psi(x)=i\Leftrightarrow D_i(x)=\max_{j\in M}D_j(x) \tag{13}$$

最后的分类精度是对测试集 V 中的每个测试数据点进行分类, 然后计算正确分类的数据点数占总点数的比例:

$$result(x)=\begin{cases}1, & \psi(x)==m(x)\\ 0, & \psi(x)!=m(x)\end{cases} \tag{14}$$

其中, $m(x)$为 x 的真实类标签, $m(x)\in M$.

$$Result=\sum_{x\in A}result(x)/num \tag{15}$$

将所有属于测试集 V 的数据点的$result(x)$相加后除以测试集 V 的数据点总个数num便可得到最终的分类精度$Result$.

(3) 改进 PM-MD

原 PM-MD 中式 (7) 采用高斯势函数将欧氏距离$\|x-x_k\|$映射到 (0, 1) 之间, 但当数据集的内在维度较高时不同样本距离经过高斯势函数转换后会非常地趋于 0, 这会弱化距离所带来的不同类的区别. 如图 2, 图 2(d)MINMAX 的距离均值较大, 大部分样本距离采用高斯势函数会得到趋于 0 的结果, 这样会使得 MINMAX 分类效果变弱, 从而影响集成效果; 这个情况在文献[7]的表 3 所给实验结果中也可以体现出来, 当数据集的内在维度较高时 (如 Ionosphere 和 Spam 等) PM-MD 的分类结果往往不是很好.

根据 PM-MD 不适用于高维数据集的情况下, 本文提出将直接采用欧氏距离计算决策表.

将 PM 进行改进:

$$H(x,x_k)=\|x-x_k\| \tag{16}$$

所对应的决策表公式:

$$\omega_j(x)=\frac{\sum_{x_k\in N_K(x):j_k=j}\|x-x_k\|}{\sum_{j\in M}\sum_{x_k\in N_K(x):j_k=j}\|x-x_k\|} \tag{17}$$

(4) 实验结果

实验中共选用 12 个 UCI 数据集[12]进行测试. 经过 10 轮十折交叉验证, 取 100 个结果的准确率均值. 12 个 UCI 数据集测试结果 (表 1) 表明:

1) 单一分类器并不适用于所有情况 (比如 NICDM 在数据集 seeds 效果最优但在数据集 mammographic_masses 效果很差), PM-MD 集成中和分类器的优劣, 在一定程度上可以获得比整体中任何单一的分类器更好的分类精度.

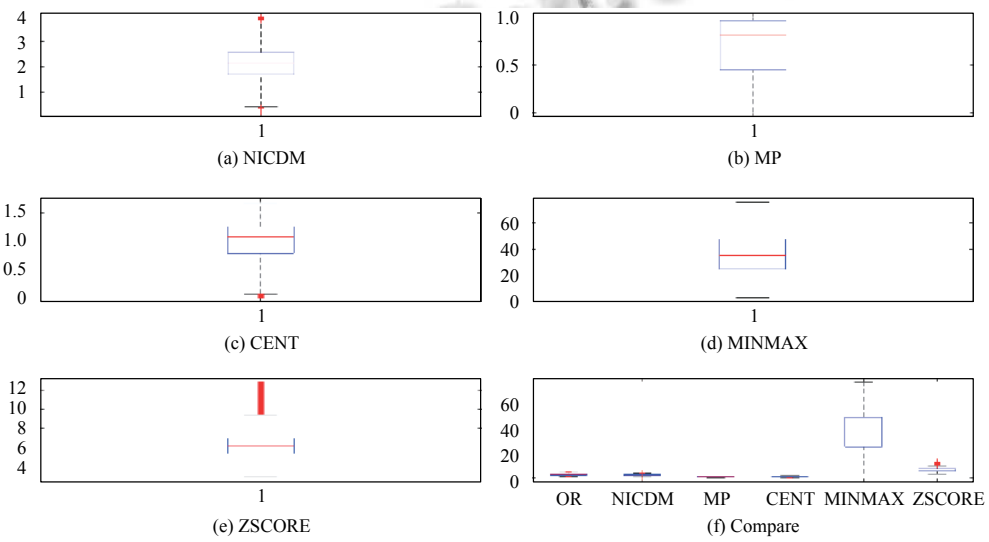

图 2 数据集 Dermatology 在各个分类器中得到的距离箱线图

2) 对于一些存在较大异常值的数据集 (比如 Pima_indians_diabetes), PM-MD 集成之后比起单一分类器有着更优的精度, 由此可见对于存在大量较大异常值的数据集, PM-MD 集成获得了比任何单一分类器要更高的分类精度.

3) 对于一些不仅存在较大异常值还存在较小异常值的数据集 (比如 wine), 集成后的分类效果明显优于大部分单一分类器分类效果, 也明显优于原始分类结果.

4) 改进后的 PM-MD 在一定程度上具有比 PM-MD 更精确的分类效果 (比如数据集 Liver), 由此可见改进后的 PM-MD 确实提升了 PM-MD 的分类效果.

5) NICDM、CNET、MINMAX、ZSCORE 的复杂度都为O(n^2), MP 的复杂度为O(n^3), 故 PM-MD 的复杂度为O(n^3).

表 1 实验结果

数据集	原始	PM-MD	改进 PM-MD	NICDM	MP	CENT	MINMAX	ZSCORE
Transfusion	0.7605	0.7633	0.7636	0.7561	0.7571	0.7665	0.7603	0.7600
Wine	0.6806	0.9633	0.9633	0.9722	0.9500	0.9600	0.9589	0.3333
processed.cleveland.data	0.5113	0.5580	0.5580	0.5613	0.5603	0.5630	0.5600	0.4927
breast-cancer-wisconsin	0.9651	0.9731	0.9729	0.9712	0.9682	0.9759	0.9669	0.9588
SPECTFincorrect.test	0.7670	0.7900	0.7895	0.7867	0.7930	0.7833	0.7822	0.3522
Pima_indians_diabetes	0.7366	0.7340	0.7347	0.7281	0.7282	0.7277	0.7281	0.6257
Haberman	0.7555	0.7342	0.7365	0.7281	0.7345	0.7300	0.7290	0.7261
Wholesale customers data	0.9041	0.8375	0.8381	0.8469	0.8353	0.8509	0.8506	0.4794
Dermatology	0.8233	0.9586	0.9589	0.9608	0.9617	0.9606	0.9581	0.7639
Liver	0.6114	0.5977	0.6017	0.5994	0.5877	0.6140	0.5874	0.5580
Seeds	0.9086	0.9148	0.9157	0.9162	0.9143	0.8976	0.9124	0.5014
mammographic_masses	0.7977	0.7805	0.7820	0.7624	0.7810	0.7759	0.7922	0.5007
PM-MD 的胜/平/负个数	7/0/5	-	2/2/8	6/0/6	7/0/5	6/0/6	9/0/3	12/0/0
改进 PM-MD 的胜/平/负个数	7/0/5	8/2/2	-	7/0/5	9/0/3	6/0/6	9/0/3	12/0/0

(5) 结论与展望

Hubness 现象是维度灾难的一个方面,hub 的出现严重降低了分类器的性能. 本文结合了五种降 Hubness 策略以减少 Hubness 的影响, 由于每种策略所适用范围的差异, 为此采用了 PM-MD 方式进行集成以扩大适用范围. 并针对 PM-MD 不适用于高维数据集的问题提出改进, 直接将欧氏距离直接用于计算决策表. 实验结果表明, PM-MD 获得了比单一分类器要高的分类精度的同时扩大了适用范围, 改进后的 PM-MD 获得了更高的分类精度. 目前研究主要关注于噪声较小的高维数据分类, 未来我们将致力于通过有效分类器集成以解决噪声较大的数据集的分类问题.

参考文献

1 Schnitzer D, Flexer A, Schedl M, *et al*. Local and global scaling reduce hubs in space. Journal of Machine Learning Research, 2012, 13(1): 2871–2902.

2 Zhai TT, He ZF. Instance selection for time series classification based on immune binary particle swarm optimization. Knowledge-Based Systems, 2013, 49: 106–115. [doi: 10.1016/j.knosys.2013.04.021]

3 Hara K, Suzuki I, Shimbo M, *et al*. Localized centering: Reducing Hubness in large-sample data. Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, TX, USA. 2015. 2645–2651.

4 Feldbauer R, Flexer A. A comprehensive empirical comparison of hubness reduction in high-dimensional spaces. Knowledge and Information Systems, 2018: 1–30. [doi: 10.1007/s10115-018-1205-y]

5 Cao H X, Stojkovic I, Obradovic Z. A robust data scaling algorithm to improve classification accuracies in biomedical data. BMC Bioinformatics, 2016, 17: 359. [doi: 10.1186/s12859-016-1236-x]

6 Feldbauer R, Flexer A. Centering versus scaling for Hubness reduction. Proceedings of the 25th International Conference on Artificial Neural Networks. Barcelona, Spain. 2016. 175–183. [doi: 10.1007/978-3-319-44778-0_21]

7 Kurzynski M, Trajdos P. On a new competence measure applied to the dynamic selection of classifiers ensemble. Proceedings of the 20th International Conference on Discovery Science. Kyoto, Japan. 2017. 93–107. [doi: 10.1007/978-3-319-67786-6_7]

8 Duda RO, Hart PE, Stork DG. Pattern Classification. New York: Wiley-Interscience, 2001.

9 Kurzynski M. On a new competence measure applied to the combining multiclassifier system. International Journal of Signal Processing Systems, 2016, 4(3): 185–191. [doi: 10.18178/ijsps.4.3.185-191]

10 Kurzynski M, Krysmann M, Trajdos P, *et al*. Multiclassifier system with hybrid learning applied to the control of bioprosthetic hand. Computers in Biology and Medicine, 2016, 69: 286–297. [doi: 10.1016/j.compbiomed.2015.04.023]

11 Meisel WS. Potential functions in mathematical pattern recognition. IEEE Transactions on Computers, 1969, C-18(10): 911–918. [doi: 10.1109/t-c.1969.222546]

12 Lichman M. UCI Machine Learning Repository. University of California. School of Information and Computer Science. http://archive.ics.uci.edu/ml, [2013-04-04].