WIKIPEDIA

最近鄰居法

维基百科，自由的百科全书

在模式识别领域中，**最近鄰居法**（**KNN**算法，又譯**K-近邻算法**）是一种用于分类和回归的無母數統計方法[1]。在这两种情况下，输入包含特征空间中的***k***个最接近的训练样本。

- 在*k-NN分类*中，输出是一个分类族群。一个对象的分类是由其邻居的"多数表决"确定的，*k*个最近邻居（*k*为正整数，通常较小）中最常见的分类决定了赋予该对象的类别。若*k* = 1，则该对象的类别直接由最近的一个节点赋予。
- 在*k-NN回归*中，输出是该对象的属性值。该值是其*k*个最近邻居的值的平均值。

最近鄰居法採用向量空間模型來分類，概念為相同類別的案例，彼此的相似度高，而可以藉由計算與已知類別案例之相似度，來評估未知類別案例可能的分類。

K-NN是一种基于实例的学习，或者是局部近似和将所有计算推迟到分类之后的惰性学习。k-近邻算法是所有的机器学习算法中最简单的之一。

无论是分类还是回归，衡量邻居的权重都非常有用，使较近邻居的权重比较远邻居的权重大。例如，一种常见的加权方案是给每个邻居权重赋值为1/ d，其中d是到邻居的距离。[註 1]

邻居都取自一组已经正确分类（在回归的情况下，指属性值正确）的对象。虽然没要求明确的训练步骤，但这也可以当作是此算法的一个训练样本集。

k-近邻算法的缺点是对数据的局部结构非常敏感。本算法与K-平均算法（另一流行的机器学习技术）没有任何关系，请勿与之混淆。

目录

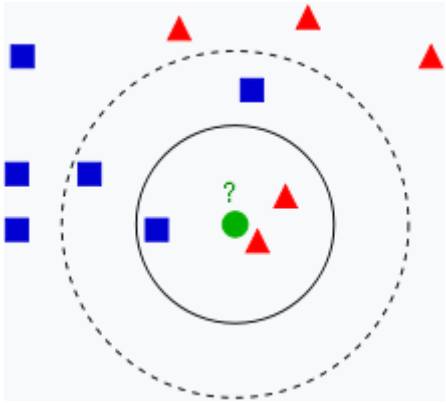
算法

训练样本是多维特征空间向量，其中每个训练样本带有一个类别标签。算法的训练阶段只包含存储的特征向量和训练样本的标签。

在分类阶段，*k*是一个用户定义的常数。一个没有类别标签的向量（查询或测试点）将被归类为最接近该点的*k*个样本点中最频繁使用的一类。

一般情况下，将欧氏距离作为距离度量，但是这是只适用于连续变量。在文本分类这种离散变量情况下，另一个度量——**重叠度量**（或海明距离）可以用来作为度量。例如对于基因表达微阵列数据，*k*-NN也与Pearson和Spearman相关系数结合起来使用。[2]通常情况下，如果运用一些特殊的算法来计算度量的话，*k*近邻分类精度可显著提高，如运用大间隔最近邻居或者邻里成分分析法。

“多数表决”分类会在类别分布偏斜时出现缺陷。也就是说，出现频率较多的样本将会主导测试点的预测结果，因为他们比较大可能出现在测试点的K邻域而测试点的属性又是通过*k*邻域内的样本计算出来的。[3]解决这个缺点的方法之一是在进行分类时将样本到*k*个近邻点的距离考虑进去。*k*近邻点中每一个的分类（对于回归问题来说，是数值）都乘以与测试点之间距离的成反比的权重。另一种克服偏斜的方式是通过数据表示形式的抽象。例如，在自组织映射（SOM）中，每个节点是相似的点的一个集群的代表（中心），而与它们在原始训练数据的密度无关。*K*-NN可以应用到SOM中。



*k*近邻算法例子。测试样本（绿色圆形）应归入要么是第一类的蓝色方形或是第二类的红色三角形。如果k=3（实线圆圈）它被分配给第二类，因为有2个三角形和只有1个正方形在内侧圆圈之内。如果k=5（虚线圆圈）它被分配到第一类（3个正方形与2个三角形在外侧圆圈之内）。

参数选择

如何选择一个最佳的K值取决于数据。一般情况下，在分类时较大的K值能够减小噪声的影响，[4] 但会使类别之间的界限变得模糊。一个较好的K值能通过各种启发式技术（见超参数优化）来获取。

噪声和非相关性特征的存在，或特徵尺度与它们的重要性不一致会使K近邻算法的准确性严重降低。对于选取和缩放特征来改善分类已经作了很多研究。一个普遍的做法是利用进化算法优化功能扩展[5]，还有一种较普遍的方法是利用训练样本的互信息进行选择特征。

在二元（两类）分类问题中，选取*k*为奇数有助于避免两个分类平票的情形。在此问题下，选取最佳经验*k*值的方法是自助法。[6]

属性

原始朴素的算法通过计算测试点到存储样本点的距离是比较容易实现的，但它属于计算密集型的，特别是当训练样本集变大时，计算量也会跟着增大。多年来，许多用来减少不必要距离评价的近邻搜索算法已经被提出来。使用一种合适的近邻搜索算法能使K近邻算法的计算变得简单许多。

近邻算法具有较强的一致性结果。随着数据趋于无限，算法保证错误率不会超过贝叶斯算法错误率的两倍[7]。对于一些K值，K近邻保证错误率不会超过贝叶斯的。

决策边界

近邻算法能用一种有效的方式隐含的计算决策边界。另外，它也可以显式的计算决策边界，以及有效率的这样做计算，使得计算复杂度是边界复杂度的函数。[8]

连续变量估计

K近邻算法也适用于连续变量估计，比如适用反距离加权平均多个K近邻点确定测试点的值。该算法的功能有：

1. 从目标区域抽样计算欧式或马氏距离；
2. 在交叉验证后的RMSE基础上选择启发式最优的K邻域；
3. 计算多元k-最近邻居的距离倒数加权平均。

發展

然而k最近鄰居法因為計算量相當的大，所以相當的耗時，Ko與Seo提出一演算法**TCFP**（**t**ext **c**ategorization using **f**eature **p**rojection），嘗試利用特徵投影法來降低與分類無關的特徵對於系統的影響，並藉此提昇系統效能，其實驗結果顯示其分類效果與k最近鄰居法相近，但其運算所需時間僅需k最近鄰居法運算時間的五十分之一。

除了針對文件分類的效率，尚有研究針對如何促進*k*最近鄰居法在文件分類方面的效果，如Han等人於2002年嘗試利用貪心法，針對文件分類實做可調整權重的k最近鄰居法**WAkNN**（**w**eighted **a**djusted **k** **n**earest **n**eighbor），以促進分類效果；而Li等人於2004年提出由於不同分類的文件本身有數量上有差異，因此也應該依照訓練集合中各種分類的文件數量，選取不同數目的最近鄰居，來參與分類。

参见

- 最邻近搜索
- 聚类分析
- 数据挖掘
- 机器学习
- 模式识别
- 预测分析
- 维数灾难
- 主成分分析
- 最小哈希

注释

1. 这个方案是一个线性插值的推广。

参考文献

1. Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician. 1992, **46** (3): 175–185. doi:10.1080/00031305.1992.10475879
2. Jaskowiak, P. A.; Campello, R. J. G. B. Comparing Correlation Coefficients as Dissimilarity Measures for Cancer Classification in Gene Expression Data. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.208.993 Brazilian Symposium on Bioinformatics (BSB 2011): 1–8 [16 October 2014]. 外部链接存在于 | website= (帮助)
3. D. Coomans; D.L. Massart. Alternative k-nearest neighbour rules in supervised pattern recognition : Part 1. k-Neares neighbour classification by using alternative voting rules. Analytica Chimica Acta. 1982, **136**: 15–27. doi:10.1016/S0003-2670 (01)95359-0. 请检查 | doi= 值 (帮助).
4. Everitt, B. S., Landau, S., Leese, M. and Stahl, D. (2011) Miscellaneous Clustering Methods, in Cluster Analysis, 5th Edition, John Wiley & Sons, Ltd, Chichester, UK.
5. Nigsch F, Bender A, van Buuren B, Tissen J, Nigsch E, Mitchell JB (2006). "Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization". Journal of Chemical Information and Modeling 46 (6): 2412–2422. doi:10.1021/ci060149f. PMID 17125183.
6. Hall P, Park BU, Samworth RJ. Choice of neighbor order in nearest-neighbor classification. Annals of Statistics. 2008, **36** (5): 2135–2152. doi:10.1214/07-AOS537.
7. Cover TM, Hart PE (1967). "Nearest neighbor pattern classification". IEEE Transactions on Information Theory 13 (1): 21–27. doi:10.1109/TIT.1967.1053964.
8. Bremner D, Demaine E, Erickson J, Iacono J, Langerman S, Morin P, Toussaint G (2005). "Output-sensitive algorithms for computing nearest-neighbor decision boundaries". Discrete and Computational Geometry 33 (4): 593–604. doi:10.1007/s00454-004-1152-0
9. E. H. Han, G. Karypis and V. Kumar, Text categorization using weight adjusted k-Nearest Neighbor classification, Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 53–65, 2001.
10. Y. J. Ko and Y. J. Seo, Text categorization using feature projections, Proceedings of the Nineteenth international conference on Computational linguistics, Volume 1, pp. 1–7, 2002.
11. B. L. Li, Q. Lu and S. W. Yu, An adaptive k-nearest neighbor text categorization strategy, ACM Transactions on Asian Language Information Processing, Volume 3 , Issue 4, pp. 215–226, 2004.
12.

13.
14.
15.
16.
17.
18.
19.
20.
21.
22.
23.
24.
25.
26.
27.
28.
29.
30.
31.
32.
33.
34.
35.
36.
37.
38.
39.
40.
41.
42.
43.
44.
45.
46.
47.
48.
49.
50.
51.
52.
53.
54.
55.
56.
57.

拓展阅读

- When Is "Nearest Neighbor" Meaningful?
- Belur V. Dasarathy (编). Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. 1991. ISBN 0-8186-8930-7.
- Shakhnarovish, Darrell, and Indyk (编). Nearest-Neighbor Methods in Learning and Vision. MIT Press. 2005. ISBN 0-262-19547-X.
- Mäkelä H Pekkarinen A. Estimation of forest stand volumes by Landsat TM imagery and stand-level field-inventory data. Forest Ecology and Management. 2004-07-26, **196** (2–3): 245–255. doi:10.1016/j.foreco.2004.02.049

- Fast k nearest neighbor search using GPU. In Proceedings of the CVPR Workshop on Computer Vision on GPU, Anchorage, Alaska, USA, June 2008. V Garcia and E. Debreuve and M. Barlaud.
- Scholarpedia article on *k*-NN
- google-all-pairs-similarity-search
