聚类算法对比研究

计算机学院 大数据与人工智能 21215122 何峙

摘 要 聚类是一种广泛应用于数据挖掘的算法。现实中的很多数据的类别标记可能是未知的，要揭示数据蕴含的性质和内在规律，则需要通过一定的数据分析技术，这一般称为“无监督学习”技术。而聚类则是其中占用重要地位的技术，其研究最多，应用最广，衍生出许许多多的算法。每种算法分为多种类别以适用在不同场景下。本文通过对这些聚类算法之间的对比阐述，让读者了解各种聚类算法的效用，以明晰其在特定场景下优缺点。

关键词 聚类，数据挖掘，无监督学习

1. 引言

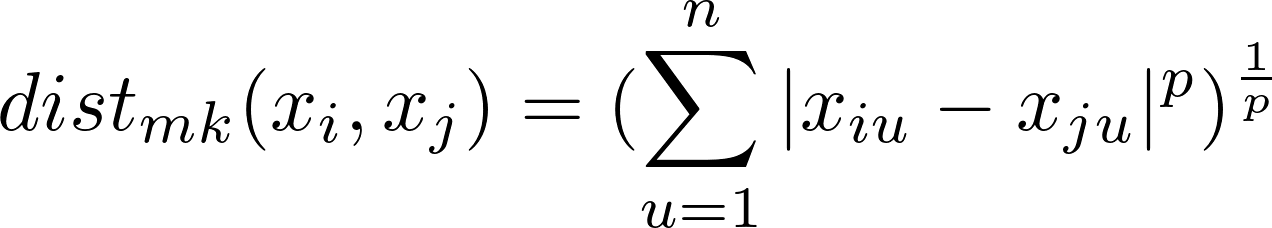
聚类是一种重要的无监督学习技术，它可以将没有类别标记的数据集划分为若干个互不相相交的子集，从而使数据样本分别位于不同的“簇”当中，这些“簇”也就是互不相交的类别，使得数据可进一步用于下游的数据分析任务。例如在一些社交应用中需要对新用户进行分类，但可能优于新用户没有填写某些兴趣爱好表单，导致系统难于定义这些新用户的用户类型，此时往往先对用户进行聚类，根据聚类结果将每个簇定义为一个类别，然后在基于这些类别训练模型，以进行下一步的内容推荐等功能应用。这时，往往会考虑：使用哪种聚类算法更适合于特定场景？还有，如何衡量这些聚类算法的性能好坏？接下来本文将从如下几个方面进行相关阐述：

* 第二章阐述聚类算法所使用的性能度量，说明数据的相似度等概念。
* 第三章列举四种不同场景下常用的聚类算法概念及其设计思想。
* 第四章在传统的聚类算法基础上，展示某些前沿的聚类算法研究。
* 第五章聚类算法总结

1. 性能度量

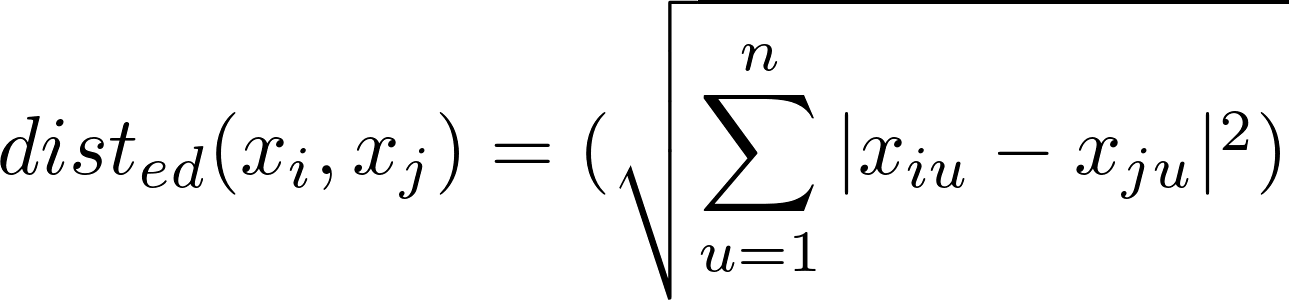
聚类算法的好坏直接影响着下游数据分析任务的优劣，那么如何衡量聚类算法的性能就相当关键。一般来说，聚类的核心性能指标是：使簇内相似度尽可能高，簇间相似度尽可能低。相似度有的文献称为“数据距离”，是一种衡量数据之间距离的度量。假设 wpsoffice 是两个样本对应的向量，以下列举的聚类中常用的相似度计算方法：

* 闵可夫斯基距离：



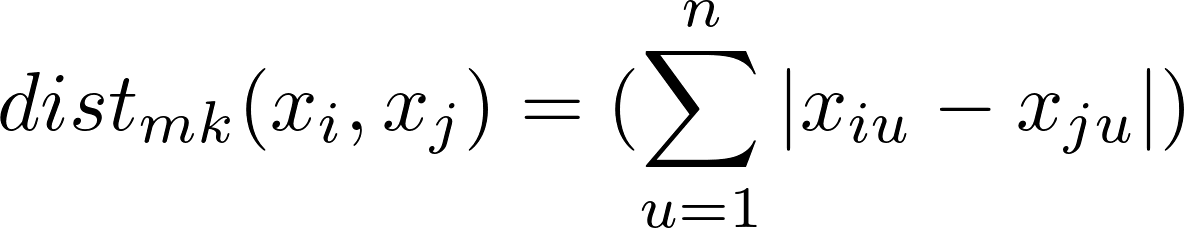
* 欧氏距离：

当闵可夫斯基距离的p=2时，即为欧氏距离：

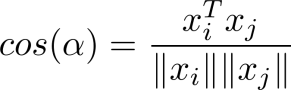


* 曼哈顿距离：

当闵可夫斯基距离的p=1时，即为曼哈顿距离：

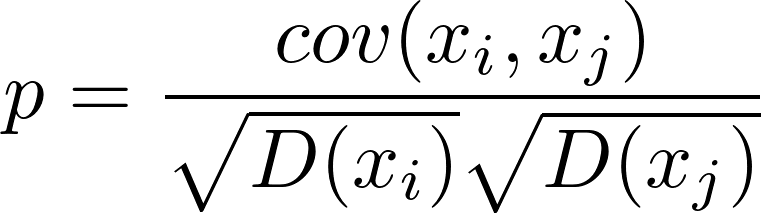


* 余弦相似度：



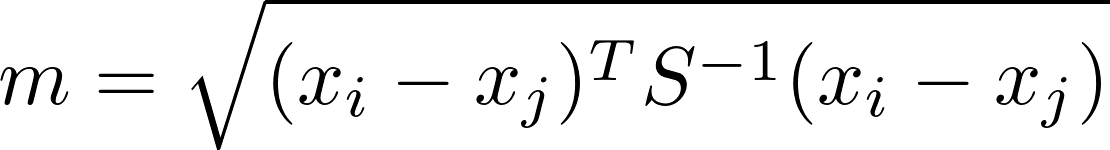
计算的结果是两个向量之间的夹角。所以值越小，可认为两个向量相似度越低。否则越大。

* 皮尔逊相关系数：



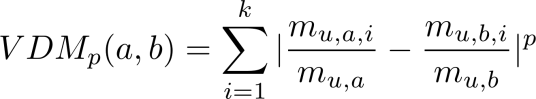
其中 wpsoffice 是两个向量的协方差，wpsoffice 是向量的方差。皮尔逊相关系数的取值范围在[-1,1]之间，值越趋于0表面向量越不相关，越趋于-1表明向量越负相关，而越趋于1明白向量越正相关。

* 马氏距离：



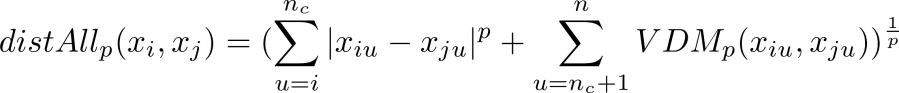
其中S为两个向量对应的协方差矩阵。

以上这些是常用于连续的属性，或者更确切的称为“有序的属性”。而对于样本中离散的、无序的属性，譬如属性的取值为学科类别={物理，化学，法律，经济}，又该如何衡量它们之间的相似度？可以考虑采用VDM数值（Value Difference Metric[hoho引用1]）衡量：



其中 /private/var/folders/4b/b2hj0zj524lbbtvttkfwrkbw0000gp/T/com.kingsoft.wpsoffice.mac/wpsoffice.akprRhwpsoffice表示在属性u上取值为a的样本数，wpsoffice 表示在第i个簇中在属性u上取值为a的样本数。

这时，可以将闵可夫斯基距离和VDM相结合，把含有有序属性和无序属性的样本相似度计算统一起来，如下计算：



有了如上这些距离度量，就可以进行性能度量。

性能度量指标又分为外部指标和内部指标两大类。外部指标是将聚类结果跟某个参考模型相比较，譬如人类专家定义的参考模型。内部指标是不利用外部参考模型而直接考虑聚类结果。

1. 外部指标

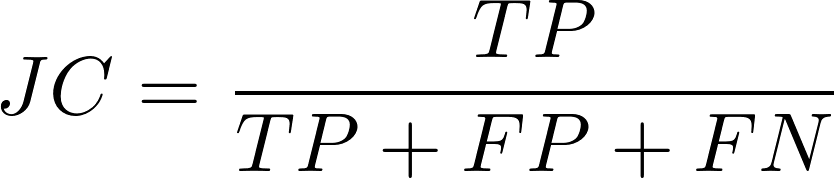
由于有外部参考模型的介入，聚类算法将各样本划分到对应簇的情况可能根外部参考模型不一致。先定义表1的混淆矩阵以方便说明问题。

|  |  |  |  |
| --- | --- | --- | --- |
|  | Acutal Value | | |
| Predicted Value |  | Positive | Negative |
| Positive | TP | FP |
| Negative | FN | TN |

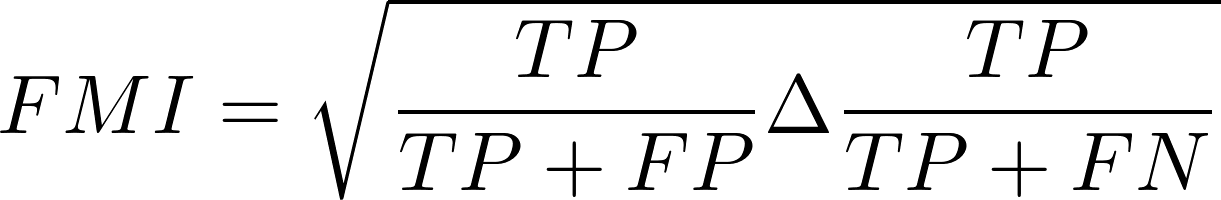
Table 1 Definition of Confusion Matrix

则有如下外部衡量指标：

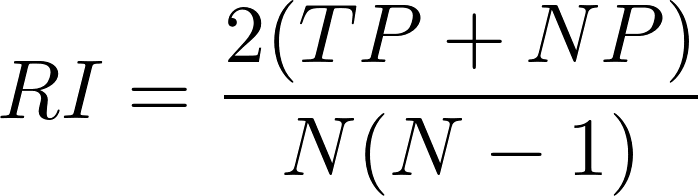
* Jaccard指数，也称Jaccard相似度：



* FM指数：



* Rand指数：



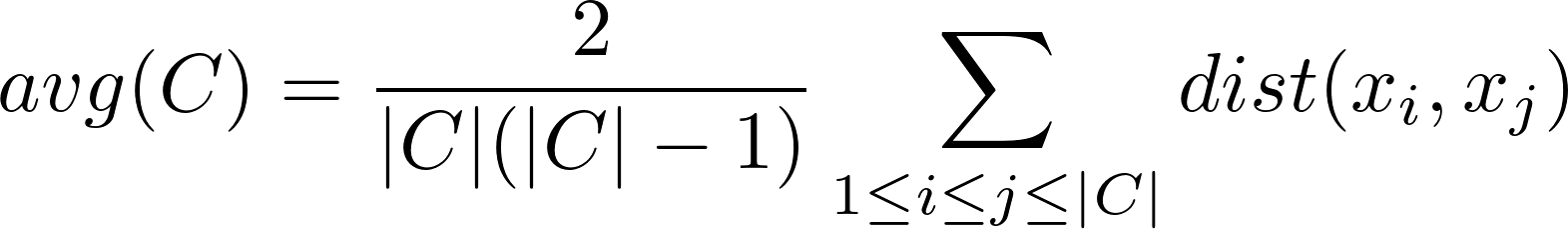
其中N为样本的数量。

显然上述性能度量的结果均在[0,1]区间，取值越大越好。

1. 内部指标

假设聚类结果的簇划分为 wpsoffice ,先定义如下计算量：

* 簇C内样本的平均距离：



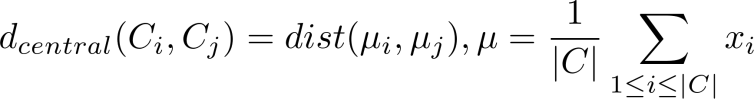
* 簇C内样本间的最远距离

wpsoffice

* 两个簇最近样本间的距离

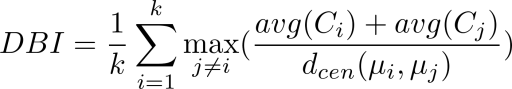


* 两个簇中心点的距离



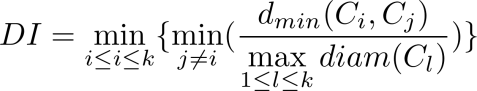
由此可导出以下常用的内部指标：

* DB指数：



显然值越小越好。

* Dunn指数



DI则相反，值越大越好。

1. 聚类算法
2. 基于划分
3. 基于密度
4. 基于层次
5. 基于网格方法
6. 前景与展望
7. 总结