

基于 Hellinger 距离的混合数据集中分类变量相似度分析

赵亮¹ 刘建辉² 王星²

(辽宁工程技术大学研究生学院 阜新 123000)¹ (辽宁工程技术大学电子与信息工程学院 葫芦岛 125000)²

摘要 分类变量的相似度分析是数据挖掘任务中的一个重要环节,现有的分类变量相似度算法中存在忽视变量差异、受不平衡分布影响严重、无法应用于混合数据集等缺点。为克服以上缺点,提出了一种基于 Hellinger 距离的分类变量相似度算法。该算法累加分类变量对应子集中不同属性变量的分布差异作为相似度,且支持混合数据集。将所提算法代入聚类算法并应用于 UCI 公共数据集,结果表明,该算法在准确度、有效性和稳定性上都有较大提高。

关键词 分类变量,相似度, f 散度, Hellinger 距离

中图分类号 TP311 文献标识码 A DOI 10.11896/j.issn.1002-137X.2016.6.055

Hellinger Distance Based Similarity Analysis for Categorical Variables in Mixture Dataset

ZHAO Liang¹ LIU Jian-Hui² WANG Xing²

(Institute of Graduate, Liaoning Technical University, Fuxin 123000, China)¹

(School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125000, China)²

Abstract Similarity analysis of categorical variables is an important part of data mining. The traditional methods have the defects of neglecting the difference between categorical variables, which are seriously affected by unbalanced dataset and can not be used in mixture dataset. To overcome the shortcomings mentioned above, this paper proposed an algorithm to measure the similarity between categorical variables based on the Hellinger distance. It accumulates the distribution differences of variables with different attributes in subsets corresponding to categorical variables as similarity variables and fits for mixture dataset. The experiments which use the derived similarity metrics in clustering algorithm and apply UCI datasets show that there is significant improvement in accuracy, validity and stability.

Keywords Categorical variables, Similarity, f-divergence, Hellinger distance

1 概述

分类变量的相似度度量在数据挖掘的许多重要应用如分类和聚类方法的设计中起着重要作用^[1,2]。但相对于连续变量在相关应用中的处理,分类变量的相似度度量更加困难,相关研究工作也较少^[3]。近年来,随着包含大量分类变量的网络数据(如电商交易数据和社交网络内容)的爆炸式增长以及相关数据分析需求的提高,分类变量相似性度量重要性也日益凸显,成为一个无法忽视的问题^[4]。

传统的分类变量相似度分析方法大体上可以分为按有监督学习和无监督学习两类。在无监督学习中,一般采用基于变量分布信息或简单累加式的度量方法^[3,5,6],其中最常见的是 Simple Matching Distance(SMD)^[6]。SMD 将相同变量之间的相似度记为 1,不同变量之间的相似度记为 0。在计算由分类变量构成的 n 维数据对象之间的相似度时,累加各个属性上的相似度并取均值作为数据对象的相似度。表 1 所列为 UCI 数据集 Balance 的一部分,该数据集由 left-weight, left-distance, right-weight, right-distance 和 balance-status 5 个数值型属性构成,分别表示天平两端砝码的重量和位置以

及最后的平衡状态,类标签 balance-status 包含表示平衡状态的 3 个分类变量 L, R 和 B 。使用 SMD 算法^[6]计算各数据项的距离时,数据项 U_1 与 U_2 以及 U_1 与 U_3 之间的相似度同为 0.25,但 U_1 与 U_2 在 balance-status 属性上同为 R ,而 U_1 与 U_3 分别为 R 和 B ;数据项 U_4 与 U_5 间的 SMD 相似度为 0,即两者毫无关系,而事实上两者同属分类 L ,以上结果的错误显而易见。引发错误的原因在于 SMD 简单地将变量之间的关系分为相同和不同,而忽视了不同变量之间相似度的差异。

表 1 Balance 数据集范例

	left-weight	left-distance	right-weight	right-distance	balance-status
U_1	1	2	1	4	R
U_2	2	1	2	4	R
U_3	3	2	2	3	B
U_4	4	3	5	2	L
U_5	5	4	4	1	L
U_6	3	5	3	5	B

有监督学习中,Stanfill 和 Waltz 提出了值差异度量(Value Difference Metric, VDM)^[7],Cost 和 Salzburg 在其基础上提出了改进值差异度量(Modified Value Difference Metric,

到稿日期:2015-05-17 返修日期:2015-09-29 本文受国家自然科学基金项目:语义 Web 模糊规则互换与推理关键技术研究(61402212)资助。
赵亮(1979-),男,博士生,主要研究方向为数据挖掘、机器学习,E-mail:tttimefighter@163.com;刘建辉(1949-),男,教授,博士生导师,主要研究方向为人工智能、计算机网络与应用;王星(1983-),男,副教授,硕士生导师,主要研究方向为智能数据与知识工程。

MVDM)^[8], Wilson 和 Martinez 提出了异构值差异度量 (Heterogeneous Value Difference Metric, HVDM)^[9]。这一类算法借鉴了无监督学习中认为同一属性中发生频率相近的变量比较相似的做法,用变量的分布以及变量与类标签的共同发生的频率计算分类变量之间的相似度。作为此类相似度算法的代表,MVDM 的计算公式为:

$$d(x, y) = \sum_{c=1}^k |p_i^x(c) - p_i^y(c)| \quad (1)$$

其中, k 代表数据集中类的数量, $p_i^x(c)$ 代表第 c 类和第 i 个属性列中 x 发生的条件概率。这类算法的缺点是无法适用于变量分布不均衡的数据集,原因在于若个别变量的条件概率值远大于其他变量的,则会掩盖其他变量间的差异。在 UCI 的数据集 Adult 中, work-class 就是这样一个分布非常不均衡的属性(见图 1)。

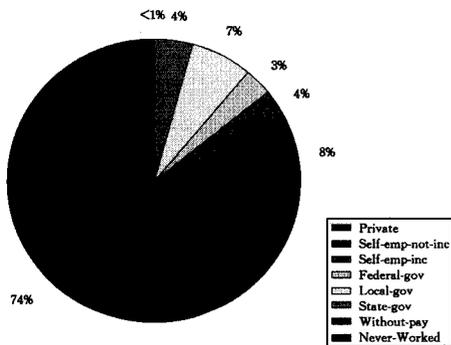


图 1 work-class 的属性分布

Adult 数据集采自美国 1994 年的收入普查数据,数据采集的目标为可能影响收入的相关信息,所以由该数据集得到的变量相似度也应该与收入相关。如图 1 所示,在 work-class 属性中,变量 Private 分布的比例最大,且远大于其他变量。运用 MVDM 算法得到的各变量与 Private 相似度(见图 2)的值集中在 1.2 附近,意味着所有职业之间在收入水平上没有明显差距,这样的结果显然不准确。

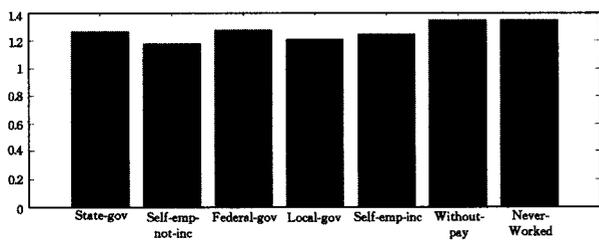


图 2 采用 MVDM 算法得到的相似度

在后续的研究中,Ahmad 和 Dey 在 MVDM 的基础上提出一种快速算法,在考虑所有属性间的相互关系的前提下形成一种精度更高的相似度指数^[10]并将其用于无监督学习。Wang C 进一步考虑了属性内部关系,提出一种耦合相似度算法^[11]。以上两个算法虽然通过加入更多的属性间关系来提高相似度的精确度,但无法从根本上解决此类算法不适用于分布不均衡数据集的缺陷。另外,以上方法均无法充分利用混合数据集提供的所有信息,甚至只能应用于分类变量数据集。国内关于分类变量度量的研究相对有限,梁吉业等人提出了一种基于粗糙集的距离度量^[12]。在聚类应用中,一些研究绕过变量之间的距离度量,通过改变数据对象间距离的计算方法来提高聚类精度^[13]。

针对已有方法存在的上述问题,本文提出一种基于 Hellinger 距离的相似度 (Hellinger Distance Similarity, HDS) 算法。该方法首先将数据集按照分类变量分为子集,计算所有子集中所有属性数据分布间的 Hellinger 距离并形成相似度向量,最后以向量各维度之和作为分类变量间的相似度。除了性能较传统算法有较大提高外,本算法还适用于由分类变量和连续变量共同构成的混合数据集。

2 基于 Hellinger 距离的分类变量相似度(HDS)算法

2.1 Hellinger 距离

在概率论中, f 散度由 Csiszar^[14], Morimoto^[15] 以及 Ali&Silvey^[16] 分别独立提出和研究,所以也被称为 Csiszar f 散度、Csiszar-Morimoto 散度或 Ali-Silvey 散度,是度量两个概率分布 P 与 Q 之间差异性的函数。设 $f(t)$ 是定义在 $t > 0$ 区间上且 $f(1) = 0$ 的凸函数, P 和 Q 是两个概率分布,则 P 与 Q 之间的 f 散度为:

$$d_f(P, Q) = \sum_x Q(x) f\left(\frac{P(x)}{Q(x)}\right) \quad (2)$$

不同的 $f(t)$ 函数使 f 散度成为不同的特例,比较常见的有 Kullback-Liebler 散度、Jensen-Shannon 散度、Pearson- χ^2 散度和 Hellinger 距离。4 种特例中只有 Hellinger 距离同时满足非负性、对称性和三角公式 3 个条件,这也是本算法在 f 散度的诸多特例中选择 Hellinger 距离作为度量方法的原因。当 $f(t) = 1 - \sqrt{t}$ 时, f 散度称为 Hellinger 距离。设 P 与 Q 为可度量空间上变量 λ 的两个分布,当 λ 为连续变量时, P 与 Q 之间 Hellinger 距离的计算公式为:

$$d_H^2(P, Q) = \frac{1}{2} \int (\sqrt{dP} - \sqrt{dQ})^2 \quad (3)$$

当 λ 为离散变量时, P 与 Q 之间 Hellinger 距离的计算公式为:

$$d_H^2(P, Q) = \frac{1}{2} \sum_{\lambda \in \Phi} (\sqrt{P(\lambda)} - \sqrt{Q(\lambda)})^2 \quad (4)$$

2.2 基于 Hellinger 距离的分类变量相似度(HDS)

算法 1

输入: 包含连续变量和分类变量的混合数据集 D

输出: 包含所有属性中同属性分类变量间的相似度

基于 Hellinger 距离的分类变量相似度函数

```
{
  For 每列属性  $A_i$  {
    if  $A_i$  中的变量为分类变量
      将数据集  $D$  根据  $A_i$  中的分类变量分为子集
      For 每对子集  $(w_1, w_2)$  对应的分布  $P$  和  $Q$  {
        For 每列其他属性  $A_j$  {
          if  $A_j$  中的变量为连续变量
            用式(3)计算子集中分布间的 Hellinger 距离
             $d_H(w_1, w_2, A_j) = \sqrt{\frac{1}{2} \int (\sqrt{dP} - \sqrt{dQ})^2}$ 
          else if  $A_j$  中的变量为分类变量
            用式(4)计算子集中分布间的 Hellinger 距离
             $d_H(w_1, w_2, A_j) = \sqrt{\frac{1}{2} \sum_{\lambda \in A_j} (\sqrt{P(\lambda)} - \sqrt{Q(\lambda)})^2}$ 
          end
        }
      }
    }
}
```

计算子集之间相似度向量各维度之和作为对应分类变量间的相似度

```

SH(w1, w2) = ∑j=1, j≠im dH(w1, w2, Aj)
}
end
}
}

```

以表 1 数据为例来说明求类标签中变量 R 和 L 的相似度的具体计算过程。首先将数据集按照分类变量分为子集, 子集 R 由数据对象 U₁ 和 U₂ 构成, 子集 L 由数据对象 U₄ 和 U₅ 构成, 表 2 给出了属性 left-weight 在两个子集中的分布。

表 2 left-weight 属性在子集 L 和子集 R 中的分布

属性变量	1	2	3	4	5
PL	0	0	0	0.5	0.5
PR	0.5	0.5	0	0	0

按照离散分布的计算公式, 两个子集在 left-weight 属性上的 Hellinger 距离为

$$\begin{aligned}
 D_H(R, L, \text{left-weight})^2 &= 0.5 * ((\sqrt{0.5} - \sqrt{0})^2 + (\sqrt{0.5} - \sqrt{0})^2 + (\sqrt{0} - \sqrt{0})^2 + \\
 &(\sqrt{0} - \sqrt{0.5})^2 + (\sqrt{0} - \sqrt{0.5})^2) \\
 &= 1
 \end{aligned}$$

在其他 3 个属性上依次计算 Hellinger 距离, 结果都为 1。L 与 R 之间在表 1 所列数据集上的相似度为

$$S_H(L, R) = 1 + 1 + 1 + 1 = 4$$

在计算 Hellinger 距离时, 需要读取两列属性, 其中一列包含被度量的分类变量, 另一列包含用来划分子集的分类变量。设数据集含有 m 列属性, n 行数据对象, 单属性列中最多包含 a 个不同的分类变量。每次计算条件概率时, 需要 m * n 步, 计算 f(t) 需要 m * a * (a - 1) / 2 步, 而整个过程重复 m

次, 则 HDS 的时间复杂度为 O(m²n + m²a²)。在大数据条件下, 数据集的数据对象个数要远大于属性与属性中变量的个数, 即 n >> m 且 n >> a, 所以近似认为 HDS 的时间复杂度对于 n 来说是线性的。

每列属性上计算所得 Hellinger 距离值越大, 代表在该属性上变量间的差异越大, 最后得到在整个数据集上的相似度的值也越大。所以在最后的结果中, 相似度值越大, 变量间的相似度实际上越低。之所以采用这种数值与实际意义相反的形式, 是为了便于将其应用在基于距离相似度的分类和聚类方法中。

3 实验分析

本实验分为两部分, 第一部分是 HDS 相似度算法在 Adult 数据集上的应用, 并将结果与之前 MVDM 算法的结果进行比较, 目的是检验 HDS 算法的准确性。第二部分用 HDS 相似度算法和 ADD^[9] 相似度算法替换 K-Modes 聚类算法中的原有 SMD 相似度算法, 并在 6 个 UCI 公共数据集上对比 3 种相似度算法的有效性及其稳定性。

3.1 HDS 的准确性

表 3 给出了由 HS 算法得到的数据集 Adult 中 work-class 属性各变量之间的相似度。从结果中可以看出, 在收入水平上, 普通的私营企业员工 (Private) 与本地以及本州的公务员 (Local-gov, State-gov) 是最相似的, 与联邦政府公务员 (Federal-gov) 稍有差距, 与私营业主 (Self-emp) 的差距更大一些, 而与无收入人群 (Without-pay, Never-worked) 的差距最大; 对于本地公务员, 其与州一级政府公务员的收入差距要小于联邦政府公务员; 私营业主间的收入差距最小。这个结果符合历年美国劳工部给出的收入水平统计, 比 MVDM 的结果显然更加准确。

表 3 HS 算法给出的 work-class 变量相似度

	private	Local-gov	Self-emp-notinc	Federal-gov	State-gov	Self-emp-inc	Without-pay	Never-worked
private	0.00	1.06	1.45	1.21	1.01	1.83	2.95	3.71
Local-gov	1.06	0.00	1.54	0.90	0.68	1.78	3.31	3.94
Self-emp-notinc	1.45	1.54	0.00	1.58	1.69	0.86	3.07	4.27
Federal-gov	1.21	0.90	1.58	0.00	0.92	1.68	3.27	4.09
State-gov	1.01	0.68	1.69	0.92	0.00	1.87	3.29	3.90
Self-emp-inc	1.83	1.78	0.86	1.68	1.87	0.00	3.33	4.54
Without-pay	2.95	3.31	3.07	3.27	3.29	3.33	0.00	4.39
Never-worked	3.71	3.94	4.27	4.09	3.90	4.54	4.39	0.00

3.2 HDS 在聚类算法中的应用

为验证 HS 的有效性, 在 K-Modes 算法中应用 SMD, ADD (Ahmad and Dey's similarity)^[10] 和 HDS 3 种不同的相似度算法, 并从精确度和稳定性两个方面比较它们在 6 个 UCI 分类数据集上的聚类结果。Vote 数据集包含 435 数据项和 16 个属性, 分为 Republican 和 Democrat 两类。Soybean 数据集包含 47 个数据项和 35 个属性, 分为 4 个分类。Zoo 数据集包含 101 数据项和 16 个属性, 分为 7 个分类。Wisconsin breast cancer 数据集包含 699 数据项和 9 个属性, 分为 Benign 和 Malignant 两类。Adult 数据集包含 48842 数据项和 14 个属性, 实验中只保留了 10 个属性, 该数据集分为两类。

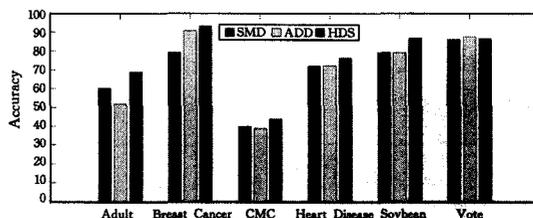


图 3 3 种相似度算法的聚类精确度比较

图 3 所示为 3 种相似度算法在 6 个 UCI 数据集上聚类结果的精确度比较。每种算法在每个数据集重复 100 次并取均值作为最终结果。在 Adult 数据集上, ADD 相似度算法得到结果的准确度最低, 这也证实了之前的观点, 单纯以分布概率

(下转第 307 页)

- [15] Taigman Y, Yang M, Ranzato M A, et al. Deepface: Closing the gap to human-level performance in face verification[C]//Proc. of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, IEEE, 2014; 1701-1708
- [16] Sun Y, Chen Y, Wang X, et al. Deep learning face representation by joint identification-verification[C]//Proc. of the Advances in Neural Information Processing Systems. 2014; 1988-1996
- [17] Phillips P J, Moon H, Rauss P J, et al. The FERET evaluation methodology for face recognition algorithms[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997; 137-143
- [18] Tan X, Triggs B. Enhanced local texture feature sets for face recognition under difficult lighting conditions[J]. IEEE TIP, 2010, 19(6): 1635-1650
- [19] Vu N S, Caplier A. Enhanced patterns of oriented edge magnitudes for face recognition and image matching[J]. IEEE TIP, 2012, 21(3): 1352-1368
- [20] Xie P, Wu X J. Modular Multilinear Principal Component Analysis and Application in Face Recognition[J]. Computer Science, 2015, 42(3): 274-279 (in Chinese)
谢佩, 吴小俊. 分块多线性主成分分析及其在人脸识别中的应用研究[J]. 计算机科学, 2015, 42(3): 274-279
- [21] Tian Hua, Pu Tian-yin. Improved ASM localization method for human facial features[J]. Journal of Chongqing University Posts and Telecommunications (Natural Science Edition), 2014, 26(1): 124-130 (in Chinese)
田华, 蒲天银. 一种改进的 ASM 人脸特征点定位方法[J]. 重庆邮电大学学报(自然科学版), 2014, 26(1): 124-130

(上接第 282 页)

的大小来区分变量间的差异在分布不均衡的数据集上效果无法让人满意。Vote 数据集也是一个比较特殊的数据集, 所有的属性都是二值属性, 由于属性变量分布太过简单, 所以 3 种相似度算法的效果几乎没有区别。

与 SMS 算法相比, HDS 算法的准确度提高最少的是在 Vote 数据集上, 只有 0.07%, 提高最多的是在 Breast cancer 数据集上, 提高了 17.85%, 6 个数据集平均提高了 9.64%。与 ADD 算法相比, HDS 算法只在 Vote 数据集上的结果降低了 1%, 在 Adult 数据集上最多提高 33%, 平均提高 10.63%。

在概率中, 方差用来度量随机变量与其数学期望的之间的偏离程度, 即数据波动幅度的大小, 方差越小, 说明算法的结果越稳定。表 4 所列为 3 种相似度算法在各数据集上聚类结果的方差, 单位为 10^{-4} 。容易看出, 除了 Soybean 数据集外, HDS 聚类结果的方差都是最小的, 说明在稳定性方面, HDS 优于 ADD 和 SMS。

表 4 聚类精度方差的比较(10^{-4})

相似度算法 \ 数据集	Adult	Breast cancer	CMC	Heart Disease	Soybean	Vote
HDS	11	71	2.1917	70	277	0.0130
SMS	63	351	8.8700	70	243	0.5206
ADD	140	92	18	164	323	0.0133

结束语 本文提出了一种基于 Hellinger 距离的分类变量相似度算法, 它适用于有监督和无监督学习且支持混合变量数据集。实验结果证明, 相比传统的 MVD, SMD, ADD 等相似度算法, 本方法不但适用范围更广, 而且在准确度、有效性和稳定性方面都有较大提高。下一步的计划是将该算法应用于其他对混合数据集支持不足的机器学习领域。

参 考 文 献

- [1] Han J, Kamber M, Pei J. Data mining: Concepts and Techniques [J]. Data Mining Concepts Models Methods & Algorithms Second Edition, 2000, 5(4): 1-18
- [2] Anderberg M R. Cluster Analysis for Applications[M]//Probability and Mathematical Statistics: A Series of Monographs and Textbooks. 1973; ibc1-ibc2
- [3] Gan G, Ma C, Wu J. Data clustering: theory, algorithms, and applications[M]//Data Clustering: theory, algorithms, and applications. Society for Industrial and Applied Mathematics, American Statistical Association, 2007; 44-51
- [4] Hanneman R A, Riddle M. Introduction to social network methods[D]. Department of Sociology, University of California Riverside, 2005
- [5] Boriah S, Chandola V, Kumar V. Similarity measures for categorical data: A comparative evaluation [J]. Proceedings of the 2008 SIAM International Conference on Data Mining, 2008, 30(2): 243-254
- [6] Huang Z. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining[C]//DMKD. 1998; 1-8
- [7] Stanfill C, Waltz D. Toward memory-based reasoning [J]. Communications of the ACM, 1986, 29(12): 1213-1228
- [8] Cost S, Salzberg S. A weighted nearest neighbor algorithm for learning with symbolic features[J]. Machine Learning, 1993, 10(1): 57-78
- [9] Wilson D R, Martinez T R. Improved heterogeneous distance functions [J]. Journal of Artificial Intelligence Research, 1997, 6: 1-34
- [10] Ahmad A, Dey L. A k-mean clustering algorithm for mixed numeric and categorical data [J]. Data & Knowledge Engineering, 2007, 63(2): 503-527
- [11] Wang C, Cao L, Wang M, et al. Coupled nominal similarity in unsupervised learning [C]//Proceedings of the 20th ACM International Conference on Information and Knowledge Management. ACM, 2011; 973-978
- [12] Liang J Y, Bai L, Cao F Y. K-Modes Clustering Algorithm Based on a New Distance Measure [J]. Journal of Computer Research and Development, 2010, 47(10): 1749-1755 (in Chinese)
梁吉业, 白亮, 曹付元. 基于新的距离度量的 K-Modes 聚类算法 [J]. 计算机研究与发展, 2010, 47(10): 1749-1755
- [13] Cao F, Liang J, Li D, et al. A dissimilarity measure for the k-Modes clustering algorithm [J]. Knowledge-Based Systems, 2012, 26: 120-127
- [14] Csizsáár I. Information-type measures of difference of probability distributions and indirect observations[M]. Studia Sci. Math. Hungar., 1967; 299-318
- [15] Morimoto T. Markov processes and the H-theorem[J]. Journal of the Physical Society of Japan, 1963, 18(3): 328-331
- [16] Ali S M, Silvey S D. A general class of coefficients of divergence of one distribution from another[J]. Journal of the Royal Statistical Society. Series B (Methodological), 1966, 28(1): 131-142