**附件2**

**论文中英文摘要**

**作者姓名：**吴俊杰

**论文题目**：考虑数据分布的K-均值聚类研究

**作者简介**：吴俊杰，男，1979年7月出生，2002年9月师从于清华大学陈剑教授，于2008年1月获博士学位。

**中 文 摘 要**

|  |
| --- |
| 商务智能通常被理解为将企业中现有的数据转化为知识，帮助企业做出明智的业务经营决策的工具。数据挖掘是商务智能的核心技术之一。实践表明，数据挖掘已经被广泛应用于客户细分和客户行为分析、目标市场定位、金融价格预测和风险管理、工作流管理、欺诈检测等商业领域，对企业的决策支持、成本管理、组织协同等提供了极大的帮助。随着信息搜索、电子商务和Web 2.0的迅猛发展，数据挖掘将为信息抽取、管理与使用发挥更大的作用。  聚类分析是数据挖掘研究的重要子领域。聚类分析为探索未知的数据结构提供帮助，并能成为一系列数据分析的起点。聚类分析已长时间在市场营销、生产监测、信息检索与分类等商业领域发挥重要作用。在聚类分析的众多算法中，K-均值算法因其简单、高效、鲁棒、数据适应性强等优点得到了非常广泛的应用。许多文献已经对K-均值算法的特点进行了深入研究，并针对海量数据、高维数据、流式数据、噪声数据等现实问题提出了许多改进方法。然而，无论从理论、算法还是实践层面，考虑到商务实践中大量数据的不均衡分布特征，仍有很多极具挑战性的问题亟待解决：   * 是否所有适于K-均值聚类的距离函数都具有统一的泛化形式？ * 是否数据的不均衡分布会对K-均值聚类带来不利影响？ * 如何在数据不均衡分布的情况下对K-均值聚类结果进行客观评价？ * 能否利用K-均值算法这样的无监督学习去提高有监督学习——如数据分布极端不平衡的稀有类分析问题——的绩效？   有鉴于此，本文围绕商务智能中广泛使用的聚类分析方法——K-均值算法，在考虑数据不均衡分布的统一框架下，做了如下工作：  **一、K-均值算法距离函数的泛化理论研究**  我们对所有适于K-均值算法的距离函数进行泛化，提出一个新的概念：K-均值距离（K-means distance）。该距离具有统一的形式，核心要素是其中的一个凸函数；通过使用不同的凸函数，可以得到一族K-均值距离。包括平方欧几里德距离、KL散度、余弦相似度等为人熟知的接近度函数，都是K-均值距离的特例。我们严格证明了，K-均值距离适用于K-均值聚类；而在一定的假设条件下，K-均值距离也是适用于K-均值聚类的唯一距离函数。提出这么一个泛化的距离函数是非常有益的，这主要体现在：1）有助于我们把握适于K-均值聚类的距离函数的共同特点，从而能够从本质上认识K-均值聚类的某些独特性质，如第三部分研究的“均匀效应”；2）有助于我们从一个较高的层面来考察以及改进传统的K-均值算法，如第二部分利用SBIL解决K-均值算法无法在稀疏数据上直接计算KL散度的问题。（相关研究发表于**ICDM 2007**[11]）  **二、SBIL算法的设计与应用研究**  我们基于K-均值距离对传统的K-均值目标函数进行了简化，并基于简化的目标函数设计了一个新颖的增量学习算法：SBIL。SBIL仍然具有K-均值算法高效运算的特点，但无需直接计算数据与簇心的K-均值距离，这对于Info-K-means（基于KL散度的K-均值算法）是至关重要的。一直以来，由于在稀疏数据上计算KL散度会出现分母为零的情况，Info-K-means被认为在文本数据上比Spherical K-means（基于余弦相似度的K-均值算法）的聚类效果差。然而，利用SBIL只需计算各簇心的某个凸函数值这一特点，我们可以巧妙地绕开KL散度的直接计算，为Info-K-means的实践应用奠定了基础。我们在大量的高维文本数据上进行了实验。实验结果表明：1）直接计算KL散度的确给Info-K-means分配数据对象带来了困难，模拟退火（annealing）和更新策略的改变没有实质性地提高Info-K-means的聚类效果；2）对数据的光滑处理可以部分地提高Info-K-means的聚类绩效，但由于数据的稀疏性被改变，而且光滑参数的设置缺乏规律性，因此在实践中很难取得令人满意的效果；3）基于SBIL的Info-K-means显示出了优良的聚类性能，其在大量数据集上的聚类效果不逊于甚至稍强于目前最好的高维数据聚类工具CLUTO实现的Spherical K-means。（相关研究发表于**KDD 2008**[9]）  **三、数据分布对K-均值算法的影响研究**  我们研究了K-均值算法与数据分布的关系。我们发现，在一定条件下（如各类中心比较接近、数据集的可聚类性不是非常好），K-均值算法倾向于生成均匀的簇，即显示所谓的“均匀效应”。我们首先从理论上论证了对于最为常用的距离函数——平方欧几里德距离，K-均值算法倾向于产生均匀的簇。其次，我们基于信息理论，对使用KL散度的K-均值算法——Info-K-means进行了研究，发现Info-K-means仍然容易导致均匀效应。最后，基于泛化的距离函数——K-均值距离，我们证明了，无论采用何种距离函数，K-均值算法都有产生均匀效应的动机。这样，我们就从理论上比较完备地论证了K-均值算法的均匀效应。接下来，我们在一系列真实数据上对K-均值算法的均匀效应进行了检验。实验结果表明，K-均值算法倾向于将数据划分为均匀的簇。这种均匀性是用统计量变异系数（Coefficient of Variation, CV）来刻画的。更进一步，我们的实验表明，任一数据集经K-均值聚类之后，得到的各簇样本量的分布符合正态分布。因此，我们可以计算K-均值聚类后各簇样本量分布的CV值的95%置信区间：[0.10, 0.72]。考虑到很多真实数据如文本数据（document data set）的各类样本数量是高度不平衡的，这个区间显然会比真实区间要小得多。因此，对于某些数据，如果其各类样本量分布的CV>0.72，使用K-均值算法进行聚类很可能带来较大的误差。（相关研究发表于**KDD 2006**[12], **Neurocomputing**[2], **IEEE Transactions on Systems, Man, and Cybernetics - Part B**[3], **Information Sciences**[4]）  **四、适于K-均值评价的聚类有效性指标研究**  很多时候，K-均值算法的均匀效应是有害的，尤其当数据集中各类样本数量显著不平衡的时候。这时，我们需要通过聚类评价指标来揭示聚类结果中的均匀效应。然而我们发现，被广泛使用的聚类评价指标——熵指标（entropy measure），在评价K-均值聚类结果时，倾向于给出过于乐观的评价，即无法揭示其中存在的均匀效应。有鉴于此，我们深入分析了熵指标存在的问题，并针对问题对熵指标进行了改进，提出用信息差异指标（Variation of Information, VI）来评价K-均值算法。理论分析和算例结果表明，信息差异指标的确是熵指标的有效改进。为了比较不同数据集上的聚类结果，我们还对信息差异指标的标准化情况进行了研究，提出了一个新的指标：NVI3。该指标是信息差异指标的一种标准化形式，与传统的标准化形式相比具有更好的评价性能。我们还对聚类评价指标进行了大量的实验研究。结果表明，熵指标在评价K-均值聚类结果时的确存在无法识别均匀效应的严重问题；与其他指标相比，我们提出的新指标NVI3的确显示出了良好的评价性能，不仅能够揭示聚类结果中存在的均匀效应，而且较好地实现了聚类结果的跨数据比较。（相关研究发表于**KDD 2009**[8], **Expert Systems with Applications**[1], **IEEE Transactions on Knowledge and Data Engineering**[5]；最新基于随机聚类评价的成果被**Information Sciences**[7]条件接收）  **五、K-均值算法在稀有类分析中的应用研究**  稀有事件如信用卡欺诈、财务困境、设备故障等是企业长期关注的管理难题，其突发性和破坏性可能给企业带来难以估量的损失。在实践需求的推动下，稀有事件预测研究迅速成为数据挖掘领域的热点问题。有鉴于此，我们将K-均值算法与有监督学习整合，提出了一个针对稀有类分析的全新框架：COG。COG的核心在于在各类数据内部而不是整个数据集上进行聚类，我们称之为“局部聚类”（local clustering）。通过局部聚类，我们可以利用K-均值算法的均匀效应使不均衡数据变得更为均衡，并且把数据中的线性不可分概念乃至复杂概念分解为相对简单的概念。为了进一步处理高度有偏数据，我们还把过抽样技术与COG结合，发展出适于稀有类分析的新技术：COG-OS。实验证明，COG无论在不均衡还是均衡数据上，都能显著提高线性分类器如SVMs的分类绩效；进一步，基于SVMs的COG-OS在稀有类分析中显示出了比目前最好的分类器更佳的分类性能。（相关研究发表于**KDD 2007**[10]，延续工作被**Data Mining and Knowledge Discovery**[6]接收）  **参考文献（[1]-[5]与附录1中代表性成果一致，加下划线作者为申请人）**   1. Junjie Wu, Hui Xiong, and Jian Chen. COG: Local Decomposition for Rare Class Analysis. *Data Mining and Knowledge Discovery Journal*, accepted as a regular paper. (SCI, IF: 2.421, 2008) 2. Junjie Wu, Hua Yuan, Hui Xiong, and Guoqing Chen. Exploiting Random Clustering for Cluster Evaluation in Information Retrieval. *Information Sciences*, accepted with minor revision. (SCI, IF: 3.095, 2008) 3. Junjie Wu, Hui Xiong, and Jian Chen. Adapting the Right Measures for K-means Clustering. in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2009)*, accepted as a regular paper. (EI待检) 4. Junjie Wu, Hui Xiong, and Jian Chen. SAIL: Summation-based Incremental Learning for Information-Theoretic Clustering. in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008)*, 740-748, 2008. (EI, AN: 20091912074375) 5. Junjie Wu, Hui Xiong, Wu Peng, and Jian Chen. Local Decomposition for Rare Class Analysis. in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007)*, 814-823, 2007. (EI, AN: 20075010974039) 6. Junjie Wu, Hui Xiong, Jian Chen, and Wenjun Zhou. A Generalization of Proximity Functions for K-means. in *Proceedings of the 2007 IEEE International Conference on Data Mining (ICDM 2007)*, 361-370, 2007. (EI, AN: 20083511480723) 7. Hui Xiong, Junjie Wu, and Jian Chen. K-means Clustering versus Validation Measures: A Data Distribution Perspective. in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006*), 779-784, 2006. (EI, AN: 20064210173657) |

关键词： 数据挖掘；聚类分析；算法设计与评价；数据特征；K-均值

**K-means Clustering: A Data Distribution Perspective**

Wu Jun Jie

# ABSTRACT

|  |
| --- |
| Recent years have witnessed the world-wide prevalence of the business intelligence concept. Business intelligence is often deemed to be a set of techniques that can transform business data into knowledge so as to provide valuable decision supports for managers in different levels. Data mining is one of the core techniques for business intelligence. In real-world applications, data mining technique has been widely used in customer segmentation and profiling, market positioning, financial prediction and risk management, work flow management, fraud detection, and so on. It is of great use for the success of enterprises in strategic decision making, cost management, and business collaboration. With the fierce explosion of the needs in information retrieval, electronic commerce and Web 2.0 today, data mining will certainly play a more and more important role in modern business life.  Cluster analysis is one of the key sub-areas of data mining researches. It helps to capture the natural structure of the data, and also serves as a useful starting point for the subsequent data analysis. Cluster analysis has long played an important role in a wide variety of fields such as marketing, production monitoring, information retrieval and document categorization, and so on. In the literature, many clustering algorithms have been proposed including the well-known K-means algorithm. K-means is a prototype-based clustering algorithm which has established itself as the most frequently and widely used clustering method due to its simplicity, robustness, high efficiency and wide adaptability. Indeed, people have investigated K-means clustering from various perspectives, e.g., the vast data volume, the high dimensionality, the dynamic data stream, and the presence of noise and outliers. However, **further study is still needed to address four important questions in the unified view of imbalanced data distributions** as follows:   * Can all the distance functions that fit K-means be generalized as a unified K-means distance? * Can the imbalanced data distribution make negative impact on the performances of K-means clustering? * How to make appropriate evaluations for K-means clustering results at the presence of imbalanced data distributions? * How to use K-means clustering to enhance the supervised learning or even the rare class analysis in the presence of extremely imbalanced data distributions?   This thesis thus focuses on K-means clustering and studies some important topics as follows:  **1. The Generalization of Proximity Functions for K-means**  A key design issue of K-means clustering is the use of proximity functions. Different choices of proximity functions for K-means can lead to quite different clustering results. In the literature, while a large amount of research work has been proposed on finding better proximity (distance) functions which can lead to a quick convergence, the common characteristics of proximity functions that fit K-means clustering (i.e. quickly converge to a solution) remain unknown. Along this line, we present a concept of “K-means distance”, which can be used to guide the choices of proximity functions that can fit K-means clustering. Indeed, we show that all proximity functions that fit K-means clustering can be generalized as K-means distance, which can be derived by a differentiable convex function. Some well-known proximity functions of K-means, such as the squared Euclidean distance, the KL-divergence, and the cosine similarity, all belongs to the family of K-means distance. A general proof of sufficient and necessary conditions for K-means distance functions is also provided. The value of K-means distance lies in two aspects: 1) It can help us better understand some intrinsic characteristics of K-means clustering, e.g., the uniform effect studied in the third part; 2) It provides to us a higher level view of K-means clustering such that we can improve it in some special data cases, e.g., the SBIL algorithm studied in the second part. (Researches published in **ICDM 2007** [11])  **2. SBIL: Problems, Algorithms, and Applications**  Information-theoretic clustering aims to exploit information theoretic measures as the clustering criteria. A common practice on this topic is so-called INFO-K-means, which performs K-means clustering with the KL-divergence as the proximity function. While expert efforts on INFO-K-means have shown promising results, a remaining challenge is to deal with high-dimensional sparse data. Indeed, it is possible that the centroids contain many zero-value features for high-dimensional sparse data. This leads to infinite KL-divergence values, which create a dilemma in assigning objects to the centroids during the iteration process of K-means. To meet this dilemma, we propose a Summation-based Incremental Learning (SBIL) method for INFO-K-means clustering. Specifically, by using an equivalent objective function, SBIL replaces the computation of the KL-divergence by the computation of the Shannon entropy. This can avoid the zero-value dilemma caused by the KL-divergence. Our experimental results on various real-world document data sets have shown that, with SBIL as a booster, the clustering performance of K-means can be significantly improved. Also, SBIL leads to quick convergence and a robust clustering performance on high-dimensional sparse data. (Researches published in **KDD 2008** [9])  **3. A Data Distribution View of K-means Clustering**  K-means is a well-known and widely used partitional clustering method. While there are considerable research efforts to characterize the key features of the K-means clustering algorithm, further investigation is needed to understand how data distributions can have impact on the performance of K-means clustering. To that end, we provide a formal and organized study of the effect of skewed data distributions on K-means clustering. Along this line, we first formally illustrate that K-means tends to produce clusters of relatively uniform size, even if input data have varied “true” cluster sizes. This variation is characterized by the Coefficient of Variation (CV) statistic. In addition, we have conducted extensive experiments on a number of real-world data sets from different application domains including text-document data sets, gene-expression data sets, and University of California, Irvine (UCI) data sets. Indeed, our experimental results show that, the 95% confidence interval for the CV values of the resultant cluster sizes is [0.10, 0.72]. This implies that for data sets with CV values of class sizes greater than 0.72, the uniform effect of K-means will take place with a very high probability, and the clustering results will be away from the “true” cluster distributions. (Researches published in **KDD 2006** [12], **Neurocomputing** [2], **IEEE Transactions on Systems, Man, and Cybernetics - Part B** [3], **Information Sciences** [4])  **4. Adapting the Right Measures for K-means Clustering**  The uniform effect of K-means can make negative impact on the cluster validity. More specifically, some well-known clustering validation measures may not have the ability to identify the uniform effect of K-means, so as to deliver unreliable scores on the clustering results. The widely used entropy measure is such an example which tends to deliver an overoptimistic score to clustering results of highly imbalanced data sets. To meet this challenge, we make an in-depth theoretic study of the problem of the entropy measure, and suggest using the Variation of Information (VI) instead. Moreover, to compare the clustering results across different data sets, we also study the normalization issues for VI, and propose a normalized VI measure: NVI3. Extensive experiments have been conducted on various real-world data sets to validate the problem of the entropy measure and the effectiveness of NVI3. The results justify our theoretic analysis, and show the tightness of the upper bound used for the normalization of VI. (Researches published in **KDD 2009** [8]**, Expert Systems with Applications** [1]**, IEEE Transactions on Knowledge and Data Engineering** [5]. A new work on random clustering-enabled cluster validation was recently accepted with minor revision by **Information Sciences** [7])  **5. COG: Local Decomposition for Rare Class Analysis**  Given its importance, the problem of predicting rare classes in large-scale multi-labeled data sets has attracted great attentions in the literature. However, the rare-class problem remains a critical challenge, because there is no natural way developed for handling imbalanced class distributions. Our research thus fills this crucial void by developing a method for Classification using lOcal clusterinG (COG). Specifically, for a data set with an imbalanced class distribution, we perform clustering within each large class and produce sub-classes with relatively balanced sizes. Then, we apply traditional supervised learning algorithms, such as Support Vector Machines (SVMs), for classification. In addition, for data sets with highly skewed class distributions, we further integrate the over-sampling technique into the COG scheme and propose the COG with over-sampling technique (COG-OS). The experimental results on various real-world data sets show that COG and COG-OS produce significantly higher prediction accuracies on rare classes than state-of-the-art methods. Furthermore, we show that COG can also improve the performances of traditional supervised learning algorithms on data sets with balanced class distributions. Finally, as a case study, we have applied COG for credit card fraud detection, and the result is very encouraging compared with state-of-the-art methods. (Researches published in **KDD2007** [10], the extended version has been accepted by **Data Mining and Knowledge Discovery** [6])  **References ([1]-[5] correspond to the ones in Appendix 1; The author underlined is the applicant)**   1. Junjie Wu, Hui Xiong, and Jian Chen. COG: Local Decomposition for Rare Class Analysis. *Data Mining and Knowledge Discovery Journal*, accepted as a regular paper. (SCI, IF: 2.421, 2008) 2. Junjie Wu, Hua Yuan, Hui Xiong, and Guoqing Chen. Exploiting Random Clustering for Cluster Evaluation in Information Retrieval. *Information Sciences*, accepted with minor revision. (SCI, IF: 3.095, 2008) 3. Junjie Wu, Hui Xiong, and Jian Chen. Adapting the Right Measures for K-means Clustering. in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2009)*, accepted as a regular paper. (EI待检) 4. Junjie Wu, Hui Xiong, and Jian Chen. SAIL: Summation-based Incremental Learning for Information-Theoretic Clustering. in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008)*, 740-748, 2008. (EI, AN: 20091912074375) 5. Junjie Wu, Hui Xiong, Wu Peng, and Jian Chen. Local Decomposition for Rare Class Analysis. in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007)*, 814-823, 2007. (EI, AN: 20075010974039) 6. Junjie Wu, Hui Xiong, Jian Chen, and Wenjun Zhou. A Generalization of Proximity Functions for K-means. in *Proceedings of the 2007 IEEE International Conference on Data Mining (ICDM 2007)*, 361-370, 2007. (EI, AN: 20083511480723) 7. Hui Xiong, Junjie Wu, and Jian Chen. K-means Clustering versus Validation Measures: A Data Distribution Perspective. in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, 779-784, 2006. (EI, AN: 20064210173657) |

Key words: Data Mining; Cluster Analysis; Algorithm Design and Evaluation; Data Factor; K-means