学校编码：10384　　分类号_____密级_____
学号：23220111153232　　UDC_____

厦门大学

硕 士 学 位 论 文

聚类成员生成以及带约束的聚类融合选择研究

Ensemble Generation Methods and Cluster Ensemble Selection with Constraints

李旋

指导教师姓名：杨 帆　助 理 教 授
专 业 名 称：模式识别与智能系统
论文提交日期：2014 年 4 月
论文答辩时间：2014 年 5 月
学位授予日期：2014 年 月

答辩委员会主席：_____
评 阅 人：_____

2014 年 5 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

聚类融合首先生成一个包含多个不同聚类成员的聚类成员集，然后将其合并为一个更准确的共识分区。学者们普遍认为对于优质的聚类融合，其聚类成员应彼此不同，同时每个聚类成员的质量也应维持在一个可接受的水平。许多算法可用于生成不同的基聚类划分。与分类集成相似，诸多研究关注不同聚类成员的生成过程，例如对不同数据子集进行聚类（随机抽样）以及对不同特征子集进行聚类（随机投影）。然而，很少有研究关注这两种不同的抽样方法在质量和差异性上的性能比较。在本文中，我们提出了一种基于随机抽样的聚类成员生成新方法，通过寻找最近邻样本的方式来填补抽样时缺失样本的类别信息（简称为RS-NN）。我们通过与基于传统 K-means 的聚类融合方法、典型的随机投影方法（随机特征子集，简称为 FS）以及另一种随机抽样方法（基于最近邻中心的随机抽样方法，简称 RS-NC）进行对比来验证该方法的有效性。实验表明，FS 总能取得更多样化的聚类成员集而 RS-NC 能取得较高的成员质量，我们提出的 RS-NN 方法则能在两者中合理地协调，在取得优异差异性的同时获得显著的性能提高。另外，为了追求更高的差异性，我们提出了一种基于 RS-NN 和 FS 的双随机抽样方法 FS-RS-NN，该方法在保证一定质量的同时取得更高的差异性，并在不增加时间代价的前提下获得可比较的甚至更优的聚类融合效果。

聚类融合已成为一个十分重要的数据分析工具，可以生成一个更强大更准确的共识聚类结果。现有研究表明，要得到较优的聚类融合结果，必须同时考虑聚类成员的质量和聚类成员之间的差异性。然而，很少有研究将其与先验背景知识结合起来。在本文中，我们首先对聚类成员的质量和差异性进行了简要的理论分析，然后提出了一个统一的框架来解决基于约束的聚类融合选择问题，其中样本之间的“必须连接”和“不能连接”约束作为给定的背景知识。我们将该问题转化为了一个联合优化问题，其中包含了基于给定约束的一致性、融合成员之间的差异性以及聚类成员的质量。该框架结合了两个决然不同却紧密相关的聚类主题：聚类融合和半监督聚类。我们研究了四种不同的聚类融合选择技术以追求高品质的聚类融合选择方案，实验充分说明了该框架的可行性和有效性。

关键词：成员生成；融合选择；聚类融合

Abstract

Cluster ensemble first generates a large library of different clustering solutions and then combines them into a more accurate consensus clustering. It is commonly accepted that for cluster ensemble to work well the member partitions should be different from each other, and meanwhile the quality of each partition should remain at an acceptable level. Many different strategies have been used to generate different base partitions for cluster ensemble. Similar to ensemble classification, many studies have been focusing on generating different partitions of the original dataset, i.e., clustering on different subsets(e.g., obtained using random sampling) or clustering in different feature spaces (e.g., obtained using random projection). However, little attention has been paid to the diversity and quality of the partitions generated using these two approaches. In this paper, we propose a novel cluster generation method based on random sampling, which uses the nearest neighbor method to fill the category information of the missing samples (abbreviated as RS-NN). We evaluate its performance in comparison with k-means ensemble, a typical random projection method (Random Feature Subset, abbreviated as FS), and another random sampling method (Random Sampling based on Nearest Centroid, abbreviated as RS-NC). Experimental results indicate that the FS method always generates more diverse partitions while RS-NC method generates high-quality partitions. Our proposed method, RS-NN, generates base partitions with a good balance between the quality and the diversity and achieves significant improvement over alternative methods. Furthermore, to introduce more diversity, we propose a dual random sampling method which combines RS-NN and FS methods. The proposed method can achieve higher diversity with good quality on most datasets.

Clustering ensemble has emerged as an important tool for data analysis, by which a more robust and accurate consensus clustering can be generated. On forming the ensembles, empirical studies have suggested that better ensembles can be obtained by simultaneously considering the quality of the ensembles and the diversity among ensemble members. However, little research efforts have been paid to incorporate

prior background knowledge. In this paper, we first provide a theoretical analysis on the effect of the diversity and quality of the ensemble members. We then propose a unified framework to solve constraint-based clustering ensemble se-lection problem, where some instance level must-link and cannot-link constraints are given as prior knowledge or background information. We formalize this problem as a combinatorial optimization problem in terms of the consistency under the constraints, the diversity among ensemble members, and the overall quality of ensembles. Our pro-posed framework brings together two distinct yet interrelated themes from clustering: ensemble clustering and semi-supervised clustering. We study four techniques for searching high-quality solutions. Experiments on benchmark datasets demonstrate the effectiveness of our framework.

Key Words: Ensemble generation; Ensemble selection; Ensemble clustering

目 录

CONTENTS

第一章 绪 论

1.1 研究背景及研究意义

近年来，对无监督数据的聚类探索和分析被广泛应用于数理统计、数据挖掘以及机器学习等领域。聚类的一个基本现状就是使用不同的聚类算法可以得到不同的聚类结果。而在聚类集成中，首先生成 N 个不同的聚类解决方案，进而结合这 N 个融合成员以得到最终的共识分区。这种集成思想被认为包含了所有融合成员的有效信息[1-9]。近年来的研究主要集中在两个方面：一是怎样产生有用且彼此不同的聚类成员[1,7,10-15]；二是如何设计对全部融合成员进行集成的共识函数[10,12,16,17]。

现有研究表明，融合成员彼此之间的差异性在聚类融合中扮演了十分重要的角色，实际上，高差异性的融合成员可以通过多种方式获得。各种聚类算法、有差别的数据表述方法和具有不同参数的同一聚类算法，均被应用于产生不同的聚类解决方案，通常我们构造成百上千个融合成员来生成最后的共识分区。除此之外，单个融合成员的质量也是聚类融合中必须考虑的另一重要因素。

传统的 K-means（KM）算法是最常见的聚类成员生成方法，随机地选取不同的初始中心，运行 N 次得到多个聚类划分。这种方法的优势在于算法操作简单，复杂度低，但对于边界模糊的数据，高维数据和非球形分布数据，该方法的结果不是很理想。另一种常见的产生聚类成员的方法则是通过随机抽样来生成有差别的数据子集[2]，该方法可以从多个方面反映数据集的复杂结构，从而较好地揭示数据的真实分布，而那些缺失的样本则通过计算其与各个聚类中心的欧氏距离，将其赋给最近的聚类中心，我们称之为基于最近邻中心的随机抽样方法（RS-NC）。由于 RS-NC 只是将少部分丢失样本赋给其最近的聚类中心，使得它与传统的 K-means（KM）算法没有显著的区别，这是因为极少数样本的加入并不会明显的改变聚类中心的位置。这使得该方法与 KM 一样，均具有较高的稳定性和较低的差异性。

在此基础上我们提出基于最近邻样本的随机抽样方法（RS-NN），首先通过对初始数据空间有放回的随机采样来构造不同的样本子集，进而对这些新的样本

子集进行 kmeans 聚类，其中初始聚类中心随机选定（k=2-10）。对于那些丢失的样本即未被抽取出来的那部分，我们寻找其最近邻的样本，用其最近邻样本的类别信息来填补这些丢失样本的类别信息。这种方法通过抽取样本的不同来构建聚类融合成员之间的差异性，最后通过共识函数将所有聚类融合成员合并成最终的聚类结果。事实证明这种方法具有显著的差异性，这也正是其相对于其他现有方法而言具有突破性意义的原因所在。

另外，与传统的聚类方法相比，聚类融合已成为一个十分重要的数据分析工具，可以生成一个更强大更准确的共识聚类结果。研究表明，要得到较优的最终集成结果，必须同时考虑每个聚类融合成员的质量以及各融合成员之间的差异性。然而，很少有学者将其与先验背景知识结合起来。因此，我们提出一个统一的框架来解决基于约束的聚类融合选择问题是至关重要的，其中样本之间的“必须连接”和“不能连接”约束作为给定的背景知识。我们将该问题转化为一个联合优化问题，其中包含了基于给定约束的一致性、融合成员之间的差异性以及聚类融合成员的质量。该框架结合了两个决然不同但却紧密相关的聚类主题：聚类融合以及半监督聚类，具有一定的研究价值和意义。

1.2 国内外研究现状

1.2.1 聚类集成研究现状

2002 年，Strehl 等在文献[10]中提出了聚类融合这一概念，并对其进行了定义：对同一数据集进行聚类得到两个或多个有差别的聚类划分，再将其进行合并以得到一个统一的共识分区，并且不使用原始数据原有的属性特征。将不同的聚类融合成员合并以得到改进的聚类结果，对于这种集成思想，近年来学者们从不同领域进行探索，如共识分类/聚类[18,19]、数据的证据积累[11]等。聚类融合方法均由以下两部分组成：聚类成员生成和共识函数设计。下面我们对聚类融合中的这两个基本步骤进行回顾，并对近年来此领域的相关研究进行简要介绍。

聚类成员生成是聚类融合的一个重点问题，旨在产生多个具有差异性的聚类结果。2002 年，Strehl 等[10]使用不同的聚类算法生成具有差异性的聚类划分。同年，Fred 等[11]通过随机选取 K-means 方法的初始聚类中心，运行多次来产生有