（第1页）各位老师，下午好！ 我叫XXX，指导老师是：XXX，我的论文题目是《基于云模型和数据场的聚类研究》，（第2页）我将从以下几个方面来介绍本文。（第3页）首先介绍本文的选题背景及意义，近年来，随着“互联网+”在各行业的不断普及和发展，数据量产生越来越多，领导者们作出决策也越来越依靠这些数据。然而当前数据具有数量大，价值密度低等特征，要想从中提取出有价值的信息，就需要进行数据挖掘，而数据挖掘的核心任务就是聚类。（第4页）经典的聚类算法或多或少存在着问题，有的算法对噪声数据敏感，噪声点的存在严重影响着聚类质量，有的算法需要用户给定经验参数，参数是否选取合理也影响着聚类质量。由此，对聚类问题进行研究是具有十分重要的价值和意义的。（第5页）本文是按照典型的聚类过程来展开研究的，依次是数据处理、算法选择、聚类结果评估。本文的研究内容主要是前面两块，数据处理，算法。（第6页）在数据处理这一块，主要研究了高维数据的特征提取。在聚类算法这一块，主要研究了数据场及聚类中心。下面我将详细展开叙述。

研究的第一点是数据处理的特征提取

受物理学中场论思想的启发，本文中将物质粒子间的相互作用及其场描述方法引入抽象的数域空间，每个对象点周围存在一个作用场，位于场内的任一对象都将受到其它对象的作用力，即数据场把任意一个数据的状态看作是数域空间中所有数据共同作用的结果，从而把个体与整体的相互作用考虑到数据分析中。

可以发现，势值随着距离的增加而减小，当超过一定距离后，势值恒接近于零。由高斯分布(正态分布)的“3规则”——在范围内包含 99.73%的数据对象，在范围之外，称为小概率事件，类似的，本文认为：给定包含个对象的数据集及其产生的数据场，对象间的相互作用范围为，即数据场中每个数据对象的影响范围是以该对象为中心、半径R等于 的邻域空间。

观察图可以得出结论，不论σ取何值，当空间中任一对象与场源的距离超过时，势值恒接近于零，也就是说场源对其范围之外的区域基本没有作用力存在。即当任意两个对象之间的距离大于时，相互之间的作用可以忽略不计。由此，改进的数据场公式如下：

改进的数据场只考虑邻域内所有对象间的相互影响，较全局而言，局部势值更能体现数据分布的紧密程度，且减少计算量，降低了时间复杂度。改进后的数据场降低了孤立点的势值，使得孤立点更容易辨认。

本节以一个形状数据集来直观的描述数据势场的分布情况。该数据集由7个簇组成，其原始数据分布如图所示。图是该数据集在不同取值的三维势场分布图，用红、黄、蓝三种渐变颜色表示势值的大小，其中，红色表示势值较大，蓝色表示势值较小。通过观察得知：在取值较小时，有许多“山峰”，表示每个点的势值都比较大；反之，取值较大时，三维势场分布接近于平面，除了中心处呈红色外，其他点都分布在蓝色区域内，表示只有中心处势值大而其他位置的势值都很小。由此可以看出，σ的取值对数据场的空间分布产生了极大的影响。

香农熵是系统不确定性的度量，熵越大，不确定性就越大。对于n个对象产生的数据场来说，如果每个对象的势值相等，则原始数据分布的不确定性越大，即具有最大的香农熵。反之，如果对象的势值很不对称，则不确定性最小，具有最小的香农熵。当势熵H取最小值时，对应的即为最优值。

聚类中心是这样的一类点：它们被很多点围绕，导致局部势值大，且与局部势值比自己大的点之间的距离也较远，通过肉眼观察得知：样本点1、6为聚类中心，周围有很多其他点；而样本点21、22距离其他点都较远，为孤立点。在图(b)决策图中，样本点1、6距离与势值都比较大；样本点21、22势值为零；其它样本点的距离都较小。由此，可以通过距离与势值确定聚类中心。

根据上述性质，本文提出一种方法自动确定最佳聚类数及聚类中心。令，将值从大到小排序得到的结果图，对于整体而言，值是相近的，而前面k个聚类中心的值大小差别较大。本文从异常检测的角度去寻找这些跳跃点，思想如下：从前往后依次计算相邻两个值的比值，直到连续两个比值都小于某个阈值时结束，则将该点前k个值对应的数据对象作为聚类中心，k为最佳聚类数目。

本文聚类算法主要可分为三个部分，分别是影响因子优化，确定聚类中心，数据点类簇的划分。并且对于高维数据集，还需进行特征提取。

第一个实验测试本文算法抗噪声干扰能力，以及对任意形状分布的处理能力；其原始分布情况如图所示，每个数据集中类簇的密度、形状以及大小均有差异，而且类簇的划分十分明显，根据肉眼观察即可得到先验知识，可以直观对每种算法聚类质量进行判断。

既可以发现任意形状分布的类簇，又能准确识别噪声数据，且整个聚类过程无需用户输入参数，聚类结果不依赖于人为因素。

分别对各个数据集进行特征提取，即利用逆向云发生器计算每个样本集中每条记录的数字特征，图展示的是三个高维数据集的聚类中心及簇的个数。

由于原始数据维度较大，聚类结果无法像二维数据集一样可视化，本节采用外部评价中的Jaccard系数和F值来评价聚类效果。本节实验中三个数据集的评价指数都达到了92%以上，表明本文中聚类算法具有良好的识别效果。同时也验证了特征提取方法的有效性。

根据整个聚类过程，本文的创新点主要有：

最后，我想谈谈这篇论文和系统存在的不足。这篇论文的写作的过程，也是我越来越认识到自己知识缺乏的过程。

盲审专家在肯定了本论文的研究之外，也提出了一些不足。主要有：