(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 102135979 A
(43)申请公布日 2011.07.27

(21)申请号 201010578479.9

(22)申请日 2010.12.08

(71)申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为基地总部办公楼

(72)发明人 王静毅 吴向阳 荀鹏

(74)专利代理机构 北京中博世达专利商标代理有限公司 11274
代理人 申健

(51)Int.Cl.
*G06F 17/30*(2006.01)

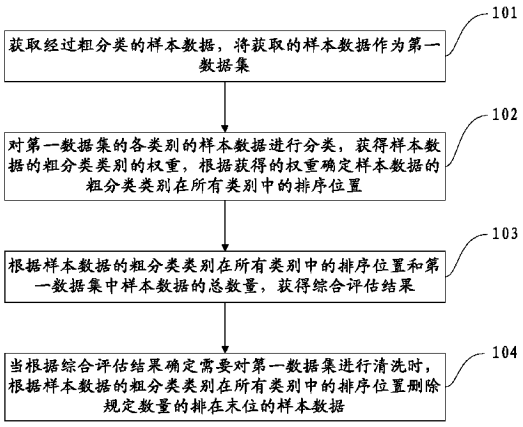
权利要求书 4 页 说明书 9 页 附图 3 页

(54)发明名称

数据清洗方法及装置

(57)摘要

本发明实施例公开了一种数据清洗方法及装置，涉及通信领域。为了能够提高数据分类的准确性，本发明提供的技术方案如下：获取经过粗分类的样本数据，将获取的样本数据作为第一数据集；对所述样本数据进行分类，获得所述样本数据的粗分类类别的权重，根据所述权重确定所述样本数据的粗分类类别在所有类别中的排序位置；根据所述样本数据的粗分类类别在所有类别中的排序位置和第一数据集中样本数据的总数量，获得综合评估结果；当根据所述综合评估结果确定需要对所述第一数据集进行清洗时，根据所述样本数据的粗分类类别在所有类别中的排序位置删除规定数量的排在末位的样本数据。本发明适用于数据分类处理。




CN 102135979 A

1. 一种数据清洗方法，其特征在于，包括：

获取经过粗分类的样本数据，将获取的样本数据作为第一数据集；

对所述样本数据进行分类，获得所述样本数据的粗分类类别的权重，根据所述权重确定所述样本数据的粗分类类别在所有类别中的排序位置；

根据所述样本数据的粗分类类别在所有类别中的排序位置和第一数据集中样本数据的总数量，获得综合评估结果；

当根据所述综合评估结果确定需要对所述第一数据集进行清洗时，根据所述样本数据的粗分类类别在所有类别中的排序位置删除规定数量的排在末位的样本数据。

2. 根据权利要求 1 所述的数据清洗方法，其特征在于，根据所述综合评估结果确定需要对所述第一数据集进行清洗包括：

当所述综合评估结果不大于第一阈值时，即为需要对所述第一数据集进行清洗，当所述综合评估结果大于第一阈值时，将所述第一数据集作为最终清洗结果。

3. 根据权利要求 1 或 2 所述的数据清洗方法，其特征在于，还包括：

将清洗后剩余的样本数据作为第二数据集；

判断所述第二数据集与第一数据集中样本数据的数量比是否大于约定比例，如果是，则对所述第二数据集继续进行清洗；如果否，则清洗失败，结束清洗。

4. 根据权利要求 1 所述的数据清洗方法，其特征在于，所述对所述样本数据进行分类包括：

将所述第一数据集中的每个粗分类类别的样本数据分成 n 组，n 为大于等于 2 的正整数；

将每个粗分类类别 n 组样本数据中的 m 组样本数据作为测试分类数据，剩余的 n-m 组样本数据作为训练样本数据，m 为大于等于 1 且小于 n 的正整数；

通过分类器根据所述训练样本数据对所述测试分类数据进行分类。

5. 根据权利要求 4 所述的数据清洗方法，其特征在于，

所述分类器包括 Bayes 分类器、KNN 分类器、SVM 分类器或类中心分类器。

6. 根据权利要求 5 所述的数据清洗方法，其特征在于，当通过 Bayes 分类器根据所述训练样本数据对所述测试分类数据进行分类时，所述样本数据的粗分类类别的权重由下述公式计算得到：

$$P(Ci/X) = P(X/Ci)*P(Ci)/P(X)$$

其中，$P(C_i) = \frac{C_i\text{类别的训练样本数}}{\text{训练样本总数}}$。

7. 根据权利要求 6 所述的数据清洗方法，其特征在于，

$$P(X/Ci) = P(x1/Ci)*P(x2/Ci)....P(xn/Ci)$$

其中，样本数据用一个 n 维特征向量，即：$X = \{x1, x2, ..., xn\}$，样本数据的粗分类类别共有 m 个类，分别用 C1，C2，...，Cm 表示。

8. 根据权利要求 7 所述的数据清洗方法，其特征在于，

当 $P(x_i|C_j) = 0$ 时，将所述 $P(x_i|C_j)$ 采用下式进行替代：

$$\frac{n_k + 1}{n + |\text{Vocabulary}|}$$

其中，n 为该类别中出现的特征的总数，$n_k$ 代表特征 $w_i$ 出现的次数，|Vocabulary| 为第一数据集中特征的总数；

所述特征为代表所属类别的关键词。

9. 根据权利要求 3 所述的数据清洗的方法，其特征在于，所述对所述第二数据集继续进行清洗的方法包括：

采用与第一数据集相同的分类方式；或，

直接采用对第一数据集进行处理时获得的所述样本数据的粗分类类别的权重和所述样本数据的粗分类类别在所有类别中的排序位置对所述第二数据集进行清洗。

10. 根据权利要求 3 或 9 所述的数据清洗的方法，其特征在于，

当采用与第一数据集相同的分类方式对第二数据集进行处理时，判断第二数据集的综合评估结果是否大于第一阈值，如果是，则将所述第二数据集作为最终清洗结果。

11. 根据权利要求 10 所述的数据清洗方法，其特征在于，当第二数据集的综合评估结果不大于第一阈值时，判断第二数据集的综合评估结果与第一数据集的综合评估结果之差是否大于第二预设阈值，如果是，则对数据集继续进行循环清洗；如果否，则判断清洗次数是否超过预设次数，如果未超过，则对第一数据集重新进行清洗；如果超过，则退出清洗，对第一数据集的清洗失败。

12. 根据权利要求 1 所述的数据清洗的方法，其特征在于，所述根据所述样本数据的粗分类类别在所有类别中的排序位置删除规定数量的排在末位的样本数据还包括：

当至少两个样本数据的粗分类类别在所有类别中的排序位置相同时，根据所述粗分类类别的权重对排序位置相同的所述至少两个样本数据进行排序。

13. 一种数据清洗装置，其特征在于，包括：

数据获取单元，用于获取经过粗分类的样本数据，将获取的样本数据作为第一数据集；

分类排序单元，用于对所述数据获取单元获取的样本数据进行分类，获得所述样本数据的粗分类类别的权重，根据所述权重确定所述样本数据的粗分类类别在所有类别中的排序位置；

综合评估单元，用于根据所述样本数据的粗分类类别在所有类别中的排序位置和第一数据集中样本数据的总数量，获得综合评估结果；

数据清洗单元，用于当根据所述综合评估单元获得的综合评估结果确定需要对所述第一数据集进行清洗时，根据所述样本数据的粗分类类别在所有类别中的排序位置删除规定数量的排在末位的样本数据。

14. 根据权利要求 13 所述的数据清洗装置，其特征在于，所述分类排序单元包括：

数据分组子单元，用于将所述第一数据集中的每个粗分类类别的样本数据分成 n 组，n 为大于等于 2 的正整数；

数据确定子单元，用于将每个粗分类类别 n 组样本数据中的 m 组样本数据作为测试分类数据，剩余的 n-m 组样本数据作为训练样本数据，m 为大于等于 1 且小于 n 的正整数；

数据分类子单元，用于通过分类器根据所述训练样本数据对所述测试分类数据进行分类；

权重获取子单元，用于获得所述样本数据的粗分类类别的权重；

数据排序子单元，用于根据所述权重确定所述样本数据的粗分类类别在所有类别中的排序位置。

15. 根据权利要求 14 所述的数据清洗装置，其特征在于，所述分类器包括 Bayes 分类器、KNN 分类器、SVM 分类器或类中心分类器。

16. 根据权利要求 15 所述的数据清洗装置，其特征在于，当通过 Bayes 分类器根据所述训练样本数据对所述测试分类数据进行分类时，所述权重获取子单元，具体用于根据公式 P(Ci/X) = P(X/Ci)*P(Ci)/P(X) 获取所述样本数据的粗分类类别的权重，其中，$P(C_i)=\frac{C_i\text{类别的训练样本数}}{\text{训练样本总数}}$，P(X/Ci) = P(x1/Ci)*P(x2/Ci)....P(xn/Ci)，样本数据用一个 n 维特征向量，即 :X = {x1, x2, ..., xn}，样本数据的粗分类类别共有 m 个类，分别用 C1, C2, ..., Cm 表示。

17. 根据权利要求 14 所述的数据清洗装置，其特征在于，所述数据排序子单元，具体用于当至少两个样本数据的粗分类类别在所有类别中的排序位置相同时，根据所述粗分类类别的权重对排序位置相同的所述至少两个样本数据进行排序。

18. 根据权利要求 13 所述的数据清洗装置，其特征在于，所述数据清洗单元包括 :

清洗判断子单元，用于当确定所述综合评估结果不大于第一阈值时，确定需要对所述样本数据进行清洗 ;当确定所述综合评估结果大于第一阈值时，将所述第一数据集作为最终清洗结果 ;

数据删除子单元，用于当所述清洗判断子单元确定需要对所述第一数据集进行清洗时，根据所述样本数据的粗分类类别在所有类别中的排序位置删除规定数量的排在末位的样本数据。

19. 根据权利要求 13-18 任一所述的数据清洗装置，其特征在于，所述数据获取单元，还用于将清洗后剩余的样本数据作为第二数据集，判断所述第二数据集与第一数据集中样本数据的数量比是否大于约定比例，如果是，则将所述第二数据集作为继续清洗的对象 ;如果否，则清洗失败，结束清洗 ;

所述数据清洗单元，还用于直接采用对第一数据集进行处理时获得的所述样本数据的粗分类类别的权重和所述样本数据的粗分类类别在所有类别中的排序位置对所述第二数据集进行清洗。

20. 根据权利要求 13-18 任一所述的数据清洗装置，其特征在于，所述数据获取单元，还用于将清洗后剩余的样本数据作为第二数据集，判断所述第二数据集与第一数据集中样本数据的数量比是否大于约定比例，如果是，则将所述第二数据集作为继续清洗的对象 ;如果否，则清洗失败，结束清洗 ;在确定第二数据集的综合评估结果不大于第一阈值，且第二数据集的综合评估结果与第一数据集的综合评估结果之差不大于第二预设阈值，且清洗次数未超过预设次数时，将第一数据集作为重新进行清洗的对象 ;在确定第二数据集的综合评估结果不大于第一阈值，且第二数据集的综合评估结果与第一数据集的综合评估结果之差不大于第二预设阈值，且清洗次数超过预设次数时，对第一数据集的清洗失败，结束清洗 ;

所述分类排序单元，还用于所述第二数据集的分类方式采用与第一数据集相同的分类方式，获得所述第二数据集的样本数据的粗分类类别的权重，根据所述权重确定所述第二

数据集的样本数据的粗分类类别在所有类别中的排序位置；

所述综合评估单元，还用于根据所述第二数据集的样本数据的粗分类类别在所有类别中的排序位置和第二数据集中样本数据的总数量，获得综合评估结果；

所述数据清洗单元，还用于在确定第二数据集的综合评估结果大于第一阈值时，将所述第二数据集作为最终清洗结果；在确定第二数据集的综合评估结果不大于第一阈值，且第二数据集的综合评估结果与第一数据集的综合评估结果之差大于第二预设阈值时，根据第二数据集的样本数据的粗分类类别在所有类别中的排序位置删除规定数量的排在末位的样本数据。

# 数据清洗方法及装置

## 技术领域

[0001] 本发明涉及通信领域，尤其涉及一种数据清洗方法及装置。

## 背景技术

[0002] 随着计算机技术和通讯技术的飞速发展，人们可以获得越来越多的数字化信息，但同时也需要投入更多的时间对信息进行组织和整理。为了减轻这种负担，人们开始研究使用计算机对数据进行自动分类。在实际应用中，互联网和文本库提供了大量已被粗分类的样本数据，但其存在数据分类错误等质量问题，因此，需要针对这些样本数据分类的正确性进行清洗。

[0003] 目前，使用如下方法对数据分类正确性进行清洗：将文本权重及其特征项权重交互迭代，直到文本权重及其特征项权重趋于稳定停止迭代，并且，利用最终的迭代结果删除低权重的文本。其中，每次迭代的具体操作如下：

[0004] $Wt^{(k+1)} = (A_{m\times n})^{T} \times Wf^{(k)}$

[0005] $Wf^{(k+1)} = A_{m\times n} \times Wt^{(k+1)}$

[0006] 其中，$Wt^{(k+1)}$ 是第 k+1 次迭代之后得到的特征项权重估计值的改进值，$Wf^{(k)}$ 和 $Wf^{(k+1)}$ 分别是第 k 次和第 k+1 次迭代之后得到的文本权重估计值的改进值，$A_{m\times n}$ 是特征项频次矩阵，m 是总的样本数，n 是特征项数。

[0007] 在实现本发明的过程中，现有技术中至少存在如下问题：在清洗多类别数据时需要对每个粗分类类别的数据逐类别进行清洗，由于缺乏类别间的对比，因此，粗分类类别的数据中可能保留类别区分有误的样本，这样会使最终的迭代结果不准确，从而降低数据分类的准确性。

## 发明内容

[0008] 本发明的实施例提供一种数据清洗方法及装置，能够提高数据分类的准确性。

[0009] 为达到上述目的，本发明的实施例采用如下技术方案：

[0010] 一种数据清洗方法，包括：

[0011] 获取经过粗分类的样本数据，将获取的样本数据作为第一数据集；

[0012] 对所述样本数据进行分类，获得所述样本数据的粗分类类别的权重，根据所述权重确定所述样本数据的粗分类类别在所有类别中的排序位置；

[0013] 根据所述样本数据的粗分类类别在所有类别中的排序位置和第一数据集中样本数据的总数量，获得综合评估结果；

[0014] 当根据所述综合评估结果确定需要对所述第一数据集进行清洗时，根据所述样本数据的粗分类类别在所有类别中的排序位置删除规定数量的排在末位的样本数据。

[0015] 一种数据清洗装置，其特征在于，包括：

[0016] 数据获取单元，用于获取经过粗分类的样本数据，将获取的样本数据作为第一数据集；

[0017] 分类排序单元，用于对所述数据获取单元获取的样本数据进行分类，获得所述样本数据的粗分类类别的权重，根据所述权重确定所述样本数据的粗分类类别在所有类别中的排序位置；

[0018] 综合评估单元，用于根据所述样本数据的粗分类类别在所有类别中的排序位置和第一数据集中样本数据的总数量，获得综合评估结果；

[0019] 数据清洗单元，用于当根据所述综合评估结果确定需要对所述第一数据集进行清洗时，根据所述样本数据的粗分类类别在所有类别中的排序位置删除规定数量的排在末位的样本数据。

[0020] 本发明实施例提供的数据清洗方法及装置，通过获取经过粗分类的样本数据，将获取的样本数据作为第一数据集，对所述第一数据集的样本数据进行分类，获得所述样本数据的粗分类类别的权重，根据所述权重确定所述样本数据的粗分类类别在所有类别中的排序位置，并根据所述样本数据的粗分类类别在所有类别中的排序位置和第一数据集中样本数据的总数量，获得综合评估结果，当根据所述综合评估结果确定需要对所述第一数据集进行清洗时，根据所述样本数据的粗分类类别在所有类别中的排序位置删除规定数量的排在末位的样本数据。因此，可以同时进行多类别数据的清洗，即，每个类别的样本数据不仅与该类别的样本数据作比较，还与所有其它类别的样本数据作比较，该类别的样本数据在经过排序清洗后，同一类别内的样本方差减小，数据分类的准确性得到提高。

## 附图说明

[0021] 为了更清楚地说明本发明实施例的技术方案，下面将对实施例描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0022] 图 1 为本发明实施例提供的一种数据清洗方法的流程示意图；

[0023] 图 2 为本发明实施例提供的另一种数据清洗方法的流程示意图；

[0024] 图 3 为本发明实施例提供的一种数据清洗装置的构成示意图。

## 具体实施方式

[0025] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0026] 为了能够提高数据分类的准确性，本发明实施例提供一种够数据清洗方法，如图 1 所示，包括：

[0027] 101、获取经过粗分类的样本数据，将获取的样本数据作为第一数据集；

[0028] 其中，所述“粗分类”是指样本数据库中录入时样本数据已经过粗略分类，例如，视频样本数据被粗分类为喜剧、悲剧、爱情剧等，上述喜剧、悲剧、爱情剧即为相应样本数据的粗分类类别。

[0029] 102、对所述第一数据集的各类别的样本数据进行分类，获得所述样本数据的粗分

类类别的权重，根据所述权重确定所述样本数据的粗分类类别在所有类别中的排序位置；

[0030] 举例而言，可以将所述第一数据集中的每个粗分类类别的样本数据分成 n 组，n 为大于等于 2 的正整数；将每个粗分类类别 n 组样本数据中的 m 组样本数据作为测试分类数据，剩余的 n-m 组样本数据作为训练样本数据，m 为大于等于 1 且小于 n 的正整数；通过分类器根据所述训练样本数据对所述测试分类数据进行分类。其中，该分类器可以为 Bayes（贝叶斯）分类器、KNN 分类器、SVM 分类器或者类中心分类器等。

[0031] 当通过 Bayes 分类器根据所述训练样本数据对所述测试分类数据进行分类时，所述样本数据的粗分类类别的权重由公式 P(Ci/X) = P(X/Ci)*P(Ci)/P(X) 计算得到。其中，$P(C_i)=\frac{C_i\text{类别的训练样本数}}{\text{训练样本总数}}$。P(X/Ci) = P(x1/Ci)*P(x2/Ci)....P(xn/Ci)，样本数据用一个 n 维特征向量，即：X = {x1, x2, ..., xn}，样本数据的粗分类类别共有 m 个类，分别用 C1, C2, ..., Cm 表示。并且，当 $P(x_i|C_j) = 0$ 时，将所述 $P(x_i|C_j)$ 采用 $\frac{n_k+1}{n+|\text{Vocabulary}|}$ 进行替代。其中，n 为该类别中出现的特征的总数，$n_k$ 代表特征 $w_i$ 出现的次数，|Vocabulary| 为第一数据集中特征的总数，而所述特征为代表所属类别的关键词。

[0032] 另外，当至少两个样本数据的粗分类类别在所有类别中的排序位置相同时，根据所述粗分类类别的权重对排序位置相同的所述至少两个样本数据进行排序。

[0033] 103、根据所述样本数据的粗分类类别在所有类别中的排序位置和第一数据集中样本数据的总数量，获得综合评估结果；

[0034] 104、当根据所述综合评估结果确定需要对所述第一数据集进行清洗时，根据所述样本数据的粗分类类别在所有类别中的排序位置删除规定数量的排在末位的样本数据。

[0035] 举例而言，在确定所述综合评估结果小于第一阈值时，确定需要对所述样本数据进行清洗。当所述综合评估结果大于第一阈值时，将所述第一数据集作为最终清洗结果。

[0036] 在对第一数据集的样本数据集进行清洗后，将清洗后剩余的样本数据作为第二数据集。判断所述第二数据集与第一数据集中样本数据的数量比是否大于约定比例，如果是，则对所述第二数据集继续进行清洗；如果否，则清洗失败，结束清洗。

[0037] 在对第二数据集继续进行清洗的过程中，可以在对第二数据集进行分类时，可以采用与第一数据集相同的分类方式，并且，当采用与第一数据集相同的分类方式对第二数据集进行处理时，判断第二数据集的综合评估结果是否大于第一阈值，如果是，则将所述第二数据集作为最终清洗结果。或者，直接采用对第一数据集进行处理时获得的所述样本数据的粗分类类别的权重和所述样本数据的粗分类类别在所有类别中的排序位置对所述第二数据集进行清洗。

[0038] 当第二数据集的综合评估结果不大于第一阈值时，判断第二数据集的综合评估结果与第一数据集的综合评估结果之差是否大于第二预设阈值，如果是，则对数据集继续进行循环清洗；如果否，则判断清洗次数是否超过预设次数，如果未超过，则对第一数据集重新进行清洗；如果超过，则退出清洗，对第一数据集的清洗失败。

[0039] 本实施例提供的数据清洗方法，通过获取经过粗分类的样本数据，将获取的样本数据作为第一数据集，对所述第一数据集的样本数据进行分类，获得所述样本数据的粗分类类别的权重，根据所述权重确定所述样本数据的粗分类类别在所有类别中的排序位置，

并根据所述样本数据的粗分类类别在所有类别中的排序位置和第一数据集中样本数据的总数量，获得综合评估结果，当根据所述综合评估结果确定需要对所述第一数据集进行清洗时，根据所述样本数据的粗分类类别在所有类别中的排序位置删除规定数量的排在末位的样本数据，并循环进行上述清洗操作直至数据满足条件完成清洗。因此，可以同时进行多类别数据的清洗，即，每个类别的样本数据不仅与该类别的样本数据作比较，还与所有其它类别的样本数据作比较，该类别的样本数据在经过排序清洗后，同一类别内的样本方差减小，数据分类的准确性得到提高。

[0040] 下面，以 Bayes 分类器对样本数据进行分类为例，对上一实施例做进一步详细具体的描述。

[0041] 如图 2 所示，本实施例数据清洗方法，包括：

[0042] 201、从原始训练数据库中按照粗分类类别分别读取经过粗分类的样本数据，将这些样本数据合并，作为数据集 A（即第一数据集）；

[0043] 其中，所述原始训练数据库中存储有已被粗分类的原始的训练样本集，并向分类器提供训练样本。例如，所述原始训练数据库可以为互联网或文本库等，进一步的样本数据可以为文本或视频等，以视频为例，在原始训练数据库中，视频样本数据被粗分类为喜剧、悲剧、爱情剧等，上述喜剧、悲剧、爱情剧即为相应样本数据的粗分类类别。

[0044] 202、将数据集 A 中每个粗分类类别的数据分别随机分成 n 组。

[0045] 例如，假设数据集 A 中存在 x 个类别的数据，分别将类别 a 分成 a.group1，a.group2，.....，a.group(n)，将类别 b 分成 b.group1，b.group2，.....，b.group(n)，.....，将类别 x 分成 x.group1，x.group2，.....，x.group(n)。

[0046] 203、在每个粗分类类别的 n 组数据中轮换确定 m 组数据为测试分类数据，并确定每个粗分类类别其余的 n-m 组数据为分类用的训练样本数据，通过 Bayes 分类器根据该训练样本数据对测试分类数据进行分类，得到样本数据的粗分类类别的权重，进一步的可以根据所述权重确定所述样本数据的粗分类类别在所有类别中的排序位置。

[0047] 在本实施例中，分类器包括 bayes 分类器，但不局限于该分类器，还可以使用其他的分类器进行分类，例如：KNN 分类器、SVM 分类器、类中心分类器等都可以用于数据清洗。所述分类结果集中的每个样本数据都会有样本号、样本数据的粗分类类别、样本数据的粗分类类别的权重、粗分类类别在所有类别中的排序位置等记录信息。其中，样本数据的粗分类类别的权重、粗分类类别在所有类别中的排序位置是进行分类后所得到的分类结果集的元素。

[0048] 例如，假设 m = 1，将 a.group1，a.group2，.....，a.group(n-1)，b.group1，b.group2，.....，b.group(n-1)，.....，x.group1，x.group2，.....，x.group(n-1) 作为训练样本数据，将 a.group(n)，b.group(n)，.....，x.group(n) 作为测试分类数据，根据这些训练样本数据通过 Bayes 分类器对测试分类数据进行分类，得到分类结果 1。

[0049] 将 a.group1，a.group2，.....，a.group(n-2)，a.group(n)，b.group1，b.group2，.....，b.group(n-2)，b.group(n)，.....，x.group1，x.group2，.....，x.group(n-2)，x.group(n) 作为训练样本数据，将 a.group(n-1)，b.group(n-1)，.....，x.group(n-1) 作为测试分类数据，根据这些训练样本数据通过 Bayes 分类器对测试分类数据进行分类，得到分类结果 2。

[0050] .....

[0051] 将 a.group2,.....,a.group(n),b.group2,.....,b.group(n),.....,x.group2,.....,x.group(n) 作为训练样本数据,将 a.group(1),b.group(1),.....,x.group(1) 作为测试分类数据,根据这些训练样本数据通过 Bayes 分类器对测试分类数据进行分类,得到分类结果 n。

[0052] 将分类结果 1,2,.....,n 合并,作为分类结果集 1。

[0053] Bayes 分类器的 Bayes 分类法具体可以为:假设样本数据用一个 n 维特征向量,即:X = {x1,x2,...,xn},样本数据总共有 m 个类,分别用 C1,C2,...,Cm 表示。给定一个未知的样本数据 X(即没有类标号),若 Bayes 分类法将未知的样本数据配给类 Ci,则一定是 P(Ci|X) > P(Cj|X),其中 j ≤ m,j ≠ i。

[0054] 根据贝叶斯定理,由于 P(X) 对于所有类为常数,最大化后验概率 P(Ci|X) 可转化为最大化先验概率 P(X|Ci)P(Ci)。如果训练数据集有许多属性和元组,各属性的取值互相独立,这样先验概率 P(x1|Ci),P(x2|Ci),...,P(xn|Ci) 可以由训样本据集通过下述公式求得。

[0055] 样本数据的粗分类类别的权重:

[0056] P(Ci/X) = P(X/Ci)*P(Ci)/P(X)

[0057] 其中,$P(C_i)=\frac{C_i\text{类别的训练样本数}}{\text{训练样本总数}}$。

[0058] x1...xn 为独立的事件,则:

[0059] P(X/Ci) = P(x1/Ci)*P(x2/Ci)....P(xn/Ci)

[0060] 由于 P(X) 对于所有的分类均常数,因此 P(Ci/X) 和 P(X/Ci)*P(Ci) 成正比,即 P(Ci/X) 的大小依赖于 P(Ci) 和 P(x1/Ci)..P(xn/Ci)。

[0061] 在实际的分类过程中,为了避免出现 $P(x_i|C_j)=0$ 的情况,对 $P(x_i|C_j)$ 采用下式进行替代:$\frac{n_k+1}{n+|\text{Vocabulary}|}$。其中 n 为该类别中出现的特征的总数,$n_k$ 代表特征 $w_i$ 出现的次数。|Vocabulary| 为第一数据集中特征的总数。

[0062] 所述特征为代表所属类别的关键词。

[0063] 用以上所述的方法求得样本数据的粗分类类别的权重后,根据求得的权重确定样本数据的粗分类类别在所有类别中的排序位置。例如,将样本数据按照其粗分类类别在所有类别中的位置进行排序,当至少两个样本数据的粗分类类别在所有类别中的排序位置相同时,根据所述粗分类类别的权重对排序位置相同的该至少两个样本数据进行排序。

[0064] 204、对分类结果集 1 进行综合评估,根据所述样本数据的粗分类类别在所有类别中的排序位置和第一数据集中样本数据的总数量,得到综合评估结果 R1。

[0065] 具体可以为,根据分类结果集中的样本数据的粗分类类别在所有类别中的排序位置和第一数据集中样本数据的总数量,获得综合评估结果 R1:

[0066]

$$R1=\frac{\text{样本数据归属粗分类类别的数量和}}{\text{数据集A的样本数据的总数量}}$$

[0067] 综合评估结果用于表示样本数据粗分类的正确率。其中,样本归属粗分类类别根

据样本数据的粗分类类别在所有类别中的排序位置确定。可以定义当粗分类类别的排序位置在预定位次之前时，则认为样本数据归属粗分类类别。例如，以视频样本数据为例，类别包括喜剧、悲剧、爱情剧、科幻剧等 10 个分类，预定位次为第 3 位，其中，样本数据的粗分类类别为喜剧，经过步骤 203 的分类计算后获得的排序位置为第 3 位，排在爱情剧、科幻剧之后，则可以确定粗分类类别在预定的第 3 位次，符合要求，认为该粗分类类别较为准确，因此，确定该样本数据归属所述粗分类类别。

[0068] 205、判断综合评估结果 R1 是否小于阈值 a（即为第一阈值），若 R1 ＞阈值 a，则确定不需要对数据集 A 进行清洗，进入步骤 206，若 R1 ＜＝阈值 a，则确定需要对数据集 A 进行清洗，进入步骤 207 ；

[0069] 其中，所述阈值 a 为预先设置好的，用于表示可接受的分类准确率，可以根据对样本数据分类准确率的要求高低进行灵活设定。

[0070] 206、退出清洗流程，将数据集 A 作为最终清洗结果，将数据集 A 存入目标数据库中。

[0071] 207、根据所述样本数据的粗分类类别在所有类别中的排序位置删除规定数量的排在末位的样本数据。将剩余的样本数据作为数据集 B（即第二数据集）。

[0072] 208、判断数据集 B 的样本数据总数占从原始训练数据库中读取的数据集 A 中的样本数据总数的比例，即数据集 B 与数据集 A 中样本数据的数量比是否大于约定比例。若数据集 B 与数据集 A 中样本数据的数量比大于约定比例，则确定数据集 B 中还有足够的样本数据，数据集 B 为可以用来对清洗效果进行评测的合格数据集，进入步骤 209。否则，确定数据集 B 中的样本数据总数过少，其为不能用来对清洗效果进行评测的不合格数据集，则进入步骤 216。

[0073] 209、将数据集 B 中每个类别的数据分别随机分成 n 组。

[0074] 本步骤的具体实现方式可参见步骤 202，在此不再赘述。

[0075] 另外，也可以采用与数据集 A 相同的分类方式对数据集 B 进行分类处理。

[0076] 210、在数据集 B 的每个类别的 n 组数据中，轮换确定 m 组数据为测试分类数据，并确定每个类别其余的 n-m 组数据为分类用的训练样本数据，通过 Bayes 分类器根据该训练样本数据对测试分类数据进行分类，得到分类结果集 2。

[0077] 本步骤的具体实现方式可参见步骤 203，在此不再赘述。

[0078] 211、对分类结果集 2 进行综合评估，得到综合评估结果 R2。

[0079] 例如，根据分类结果集 2 获取数据集 B 的样本数据归属粗分类类别的数量，并将综合评估结果 R2 定义为样本数据归属粗分类类别的概率。

[0080] 212、判断综合评估结果 R2 是否小于阈值 a，若 R2 ＞阈值 a，则确定不需要对数据集 B 进行清洗，进入步骤 213，若 R2 ＜阈值 a，则确定需要对数据集 B 进行清洗，进入步骤 214 ；

[0081] 213、退出清洗流程，将数据集 B 作为最终清洗结果，将数据集 B 存入目标数据库中。

[0082] 214、判断综合评估结果 R2 和 R1 之差是否大于阈值 b（即为第二阈值）。若 R2-R1 ＜＝阈值 b，则数据集 B 的分类效果没有提高，进入步骤 215 中。若 R2-R1 ＞阈值 b，则确定 B 的分类效果有提高，则继续对数据集 B 进行清洗，将数据集 B 作为数据集 A，返回步骤 201

进行清洗处理。

[0083] 另外，也可以在确定需要继续对数据集 B 进行清洗时，直接从当前的分类结果中获知粗分类类别在所有类别中的排序位置和粗分类类别的权重，然后根据粗分类类别在所有类别中的排序位置对样本数据进行排序，并且，在至少两个样本数据的粗分类类别在所有类别中的排序位置相同时，根据粗分类类别的权重对这些样本数据进行排序，删除规定数量的排在末位的样本数据。

[0084] 215、判断对数据集 A 中的样本数据进行清洗的总次数是否超过规定次数 K。若清洗的总次数超过规定次数 K，则确定对数据集 A 的样本数据已经进行过多次清洗，但每次清洗后的数据集的分类效果都没有提高，进入步骤 216。若清洗的总次数未超过规定次数 K，则可能由于对 A 的随机分组不当造成，对数据集 A 重新开始清洗流程。

[0085] 216、退出清洗流程，对数据集 A 的清洗操作失败。

[0086] 在本实施例中，配置了三个清洗判断条件，即判断综合评估结果是否小于规定的阈值，判断当前的数据集的样本数据总数占从原始训练数据库中读取的样本数据总数的比例是否大于约定比例，判断当前计算出的综合评估结果和前一次得到的综合评估结果之差是否大于规定的阈值。当然，也可以继续增加或者替换清洗判断条件。这些清洗判断条件可以通过配置文件进行配置。

[0087] 本实施例提供的数据清洗方法，通过获取经过粗分类的样本数据，将获取的样本数据作为第一数据集，对所述第一数据集的样本数据进行分类，获得所述样本数据的粗分类类别的权重，根据所述权重确定所述样本数据的粗分类类别在所有类别中的排序位置，并根据所述样本数据的粗分类类别在所有类别中的排序位置和第一数据集中样本数据的总数量，获得综合评估结果，当根据所述综合评估结果确定需要对所述第一数据集进行清洗时，根据所述样本数据的粗分类类别在所有类别中的排序位置删除规定数量的排在末位的样本数据，并循环进行上述清洗操作直至数据满足条件完成清洗。因此，可以同时进行多类别数据的清洗，即，每个类别的样本数据不仅与该类别的样本数据作比较，还与所有其它类别的样本数据作比较，该类别的样本数据在经过排序清洗后，同一类别内的样本方差减小，数据分类的准确性得到提高。并且，通过在迭代过程中逐步从读取的数据集中删除不符合清洗判断规则的文本，对训练数据进行清洗，进而可以提高数据特征提取的准确性，从而可以进一步提高数据分类的准确性。

[0088] 与上述方法相对应地，本发明实施例还提供了一种数据清洗装置，如图 3 所示，包括：

[0089] 数据获取单元 301，获取经过粗分类的样本数据，将获取的样本数据作为第一数据集；

[0090] 分类排序单元 302，用于对所述数据获取单元 301 获取的样本数据进行分类，获得所述样本数据的粗分类类别的权重，根据所述权重确定所述样本数据的粗分类类别在所有类别中的排序位置；

[0091] 综合评估单元 303，用于根据所述样本数据的粗分类类别在所有类别中的排序位置和第一数据集中样本数据的总数量，获得综合评估结果；

[0092] 数据清洗单元 304，用于当根据所述综合评估单元 303 获得的综合评估结果确定需要对所述第一数据集进行清洗时，根据所述样本数据的粗分类类别在所有类别中的排序

位置删除规定数量的排在末位的样本数据。

[0093] 进一步地，所述分类排序单元 302 具体包括：

[0094] 数据分组子单元，用于将所述第一数据集中的每个粗分类类别的样本数据分成 n 组，n 为大于等于 2 的正整数；

[0095] 数据确定子单元，用于将每个粗分类类别 n 组样本数据中的 m 组样本数据作为测试分类数据，剩余的 n-m 组样本数据作为训练样本数据，m 为大于等于 1 且小于 n 的正整数；

[0096] 数据分类子单元，用于通过分类器根据所述训练样本数据对所述测试分类数据进行分类；

[0097] 权重获取子单元，用于获得所述样本数据的粗分类类别的权重；

[0098] 数据排序子单元，用于根据所述权重确定所述样本数据的粗分类类别在所有类别中的排序位置。

[0099] 进一步地，所述分类器包括 Bayes 分类器、KNN 分类器、SVM 分类器或类中心分类器。

[0100] 进一步地，当通过 Bayes 分类器根据所述训练样本数据对所述测试分类数据进行分类时，所述权重获取子单元，具体用于根据公式 P(Ci/X) = P(X/Ci)*P(Ci)/P(X) 获取所述样本数据的粗分类类别的权重，其中，$P(C_i)=\frac{C_i\text{类别的训练样本数}}{\text{训练样本总数}}$，P(X/Ci) = P(x1/Ci)*P(x2/Ci)....P(xn/Ci)，样本数据用一个 n 维特征向量，即：X = {x1, x2, ..., xn}，样本数据的粗分类类别共有 m 个类，分别用 C1, C2, ..., Cm 表示。

[0101] 进一步地，所述数据排序子单元，具体用于当至少两个样本数据的粗分类类别在所有类别中的排序位置相同时，根据所述粗分类类别的权重对排序位置相同的所述至少两个样本数据进行排序。

[0102] 进一步地，所述数据清洗单元 304 包括：

[0103] 清洗判断子单元，用于当确定所述综合评估结果不大于第一阈值时，确定需要对所述样本数据进行清洗；当确定所述综合评估结果大于第一阈值时，将所述第一数据集作为最终清洗结果；

[0104] 数据删除子单元，用于当所述清洗判断子单元确定需要对所述第一数据集进行清洗时，根据所述样本数据的粗分类类别在所有类别中的排序位置删除规定数量的排在末位的样本数据。

[0105] 进一步地，所述数据获取单元，还用于将清洗后剩余的样本数据作为第二数据集，判断所述第二数据集与第一数据集中样本数据的数量比是否大于约定比例，如果是，则将所述第二数据集作为继续清洗的对象；如果否，则清洗失败，结束清洗；

[0106] 所述数据清洗单元，还用于直接采用对第一数据集进行处理时获得的所述样本数据的粗分类类别的权重和所述样本数据的粗分类类别在所有类别中的排序位置对所述第二数据集进行清洗。

[0107] 进一步地，所述数据获取单元，还用于将清洗后剩余的样本数据作为第二数据集，判断所述第二数据集与第一数据集中样本数据的数量比是否大于约定比例，如果是，则将所述第二数据集作为继续清洗的对象；如果否，则清洗失败，结束清洗；在确定第二数据集的综合评估结果不大于第一阈值，且第二数据集的综合评估结果与第一数据集的综合评估

结果之差不大于第二预设阈值，且清洗次数未超过预设次数时，将第一数据集作为重新进行清洗的对象；在确定第二数据集的综合评估结果不大于第一阈值，且第二数据集的综合评估结果与第一数据集的综合评估结果之差不大于第二预设阈值，且清洗次数超过预设次数时，对第一数据集的清洗失败，结束清洗；

[0108] 所述分类排序单元，还用于所述第二数据集的分类方式采用与第一数据集相同的分类方式，获得所述第二数据集的样本数据的粗分类类别的权重，根据所述权重确定所述第二数据集的样本数据的粗分类类别在所有类别中的排序位置；

[0109] 所述综合评估单元，还用于根据所述第二数据集的样本数据的粗分类类别在所有类别中的排序位置和第二数据集中样本数据的总数量，获得综合评估结果；

[0110] 所述数据清洗单元，还用于在确定第二数据集的综合评估结果大于第一阈值时，将所述第二数据集作为最终清洗结果；在确定第二数据集的综合评估结果不大于第一阈值，且第二数据集的综合评估结果与第一数据集的综合评估结果之差大于第二预设阈值时，根据第二数据集的样本数据的粗分类类别在所有类别中的排序位置删除规定数量的排在末位的样本数据。

[0111] 本实施例数据清洗装置的工作方法可参考图 1 和图 2 所示的实施例。

[0112] 本实施例提供的数据清洗装置，通过获取经过粗分类的样本数据，将获取的样本数据作为第一数据集，对所述第一数据集的样本数据进行分类，获得所述样本数据的粗分类类别的权重，根据所述权重确定所述样本数据的粗分类类别在所有类别中的排序位置，并根据所述样本数据的粗分类类别在所有类别中的排序位置和第一数据集中样本数据的总数量，获得综合评估结果，当根据所述综合评估结果确定需要对所述第一数据集进行清洗时，根据所述样本数据的粗分类类别在所有类别中的排序位置删除规定数量的排在末位的样本数据，并循环进行上述清洗操作直至数据满足条件完成清洗。因此，可以同时进行多类别数据的清洗，即，每个类别的样本数据不仅与该类别的样本数据作比较，还与所有其它类别的样本数据作比较，该类别的样本数据在经过排序清洗后，同一类别内的样本方差减小，数据分类的准确性得到提高。

[0113] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，所述的程序可存储于一计算机可读取存储介质中，该程序在执行时，可包括如上述各方法的实施例的流程。其中，所述的存储介质可为磁碟、光盘、只读存储记忆体 (Read-Only Memory，ROM) 或随机存储记忆体 (Random Access Memory，RAM) 等。

[0114] 以上所述，仅为本发明的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应所述以权利要求的保护范围为准。

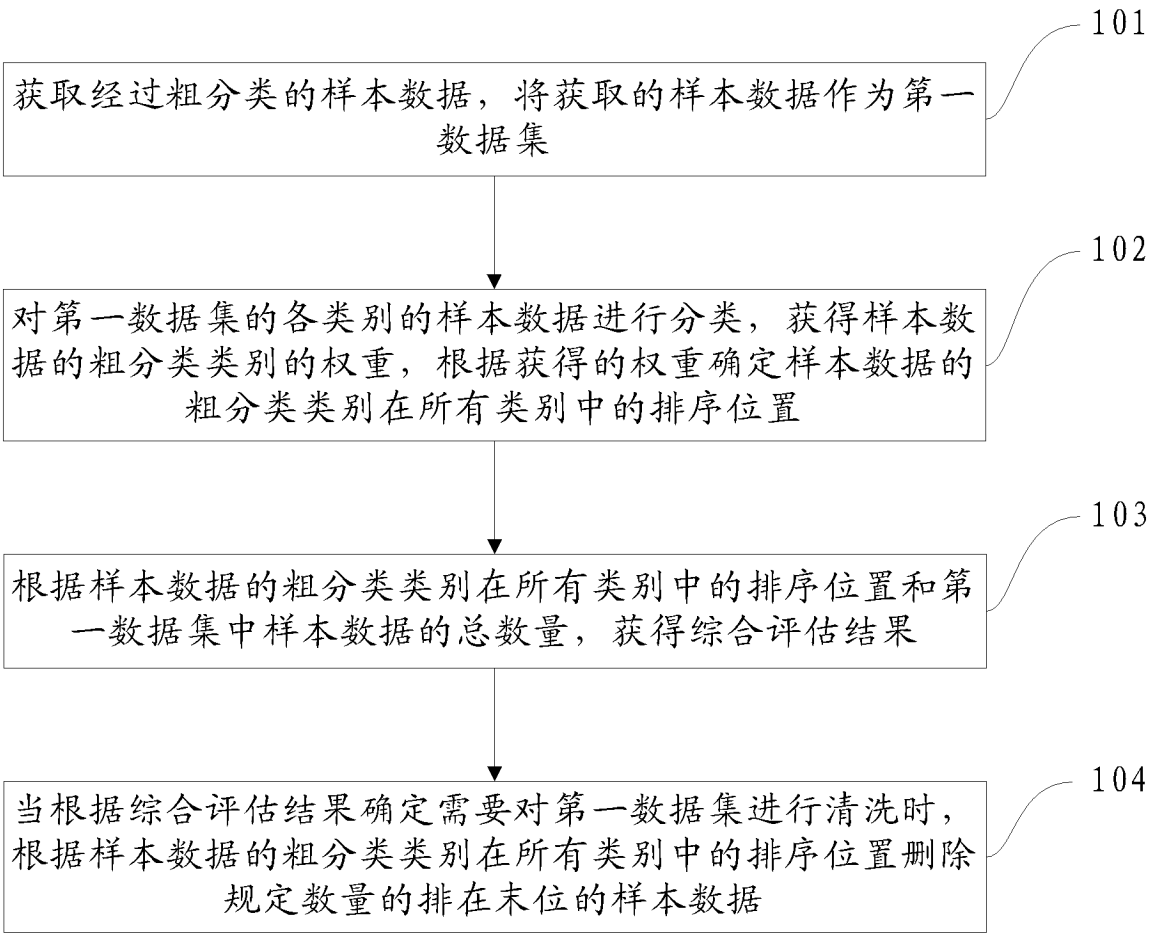

101
获取经过粗分类的样本数据，将获取的样本数据作为第一数据集
102
对第一数据集的各类别的样本数据进行分类，获得样本数据的粗分类类别的权重，根据获得的权重确定样本数据的粗分类类别在所有类别中的排序位置
103
根据样本数据的粗分类类别在所有类别中的排序位置和第一数据集中样本数据的总数量，获得综合评估结果
104
当根据综合评估结果确定需要对第一数据集进行清洗时，根据样本数据的粗分类类别在所有类别中的排序位置删除规定数量的排在末位的样本数据

图 1

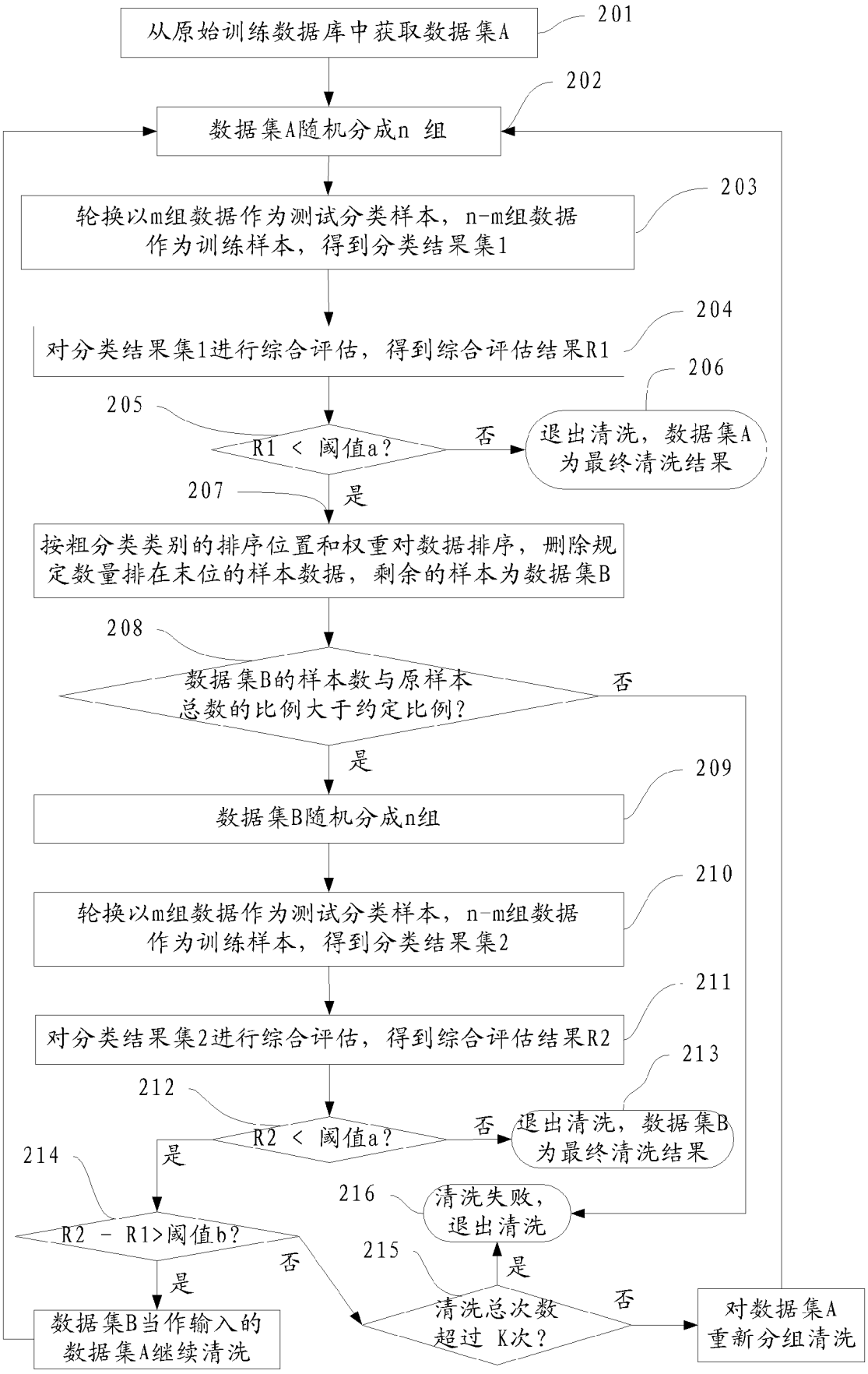

从原始训练数据库中获取数据集A
201
数据集A随机分成n 组
202
轮换以m组数据作为测试分类样本，n-m组数据作为训练样本，得到分类结果集1
203
对分类结果集1进行综合评估，得到综合评估结果R1
204
R1 < 阈值a?
205
否
退出清洗，数据集A为最终清洗结果
206
是
207
按粗分类类别的排序位置和权重对数据排序，删除规定数量排在末位的样本数据，剩余的样本为数据集B
数据集B的样本数与原样本总数的比例大于约定比例?
208
否
是
数据集B随机分成n组
209
轮换以m组数据作为测试分类样本，n-m组数据作为训练样本，得到分类结果集2
210
对分类结果集2进行综合评估，得到综合评估结果R2
211
R2 < 阈值a?
212
否
退出清洗，数据集B为最终清洗结果
213
是
214
R2 - R1>阈值b?
否
是
数据集B当作输入的数据集A继续清洗
清洗失败，退出清洗
216
清洗总次数超过 K次?
215
是
否
对数据集A重新分组清洗

图 2

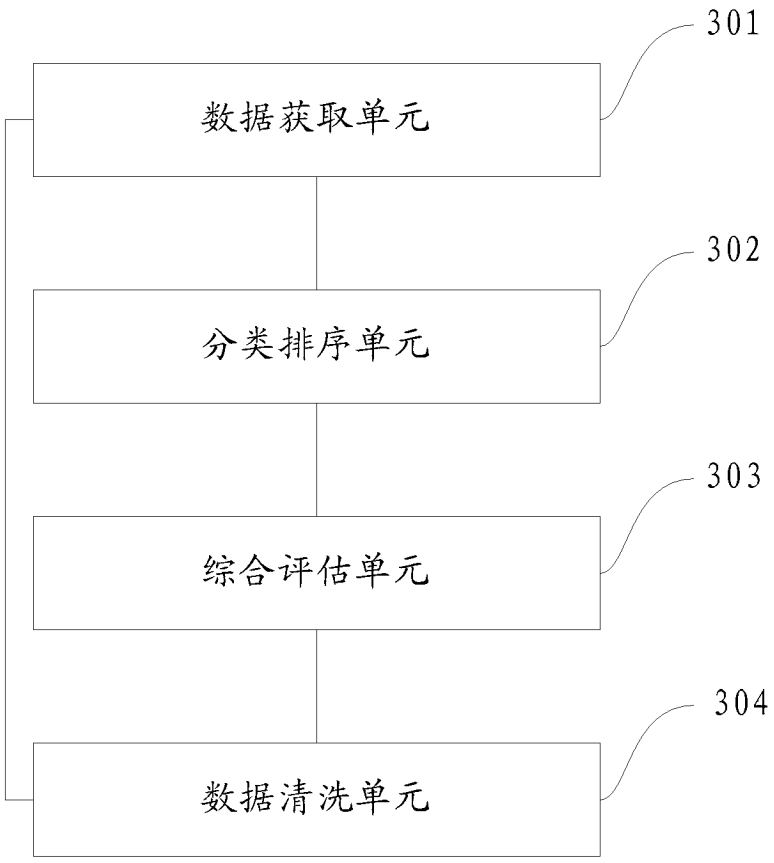



图 3