大数据下数据预处理方法研究

孔 钦,叶长青,孙 赟
(南京大学,江苏 南京 210089)

摘 要: 大数据时代下,数据类型和组织模式多样化、关联关系繁杂、质量良莠不齐等内在的复杂性使得数据的感知、表达、理解和计算等多个环节面临着巨大的挑战。数据预处理是数据分析、挖掘前一个非常重要的数据准备工作。一方面它可以保证挖掘数据的正确性和有效性,另一方面通过对数据格式和内容的调整,使数据更符合挖掘的需要。文中分析了预处理过程中的主要任务,总结了目前针对各类“脏数据”的几种常用的处理方法,重点阐述了数据在清洗、集成、变换和归约过程中的常用算法。通过各种预处理方法,清除冗余数据,纠正错误数据,完善残缺数据,甄选出必需的数据进行集成,使得数据信息精练化、数据格式一致化和数据存储集中化。在最精确、最可靠的最小数据集合上进行数据挖掘,大大减少了系统挖掘的开销,提高了知识发现的准确性、有效性和实用性。

关键词: 大数据; 预处理; 脏数据; 研究
中图分类号: TP301 **文献标识码:** A **文章编号:** 1673-629X(2018)05-0001-04
doi: 10.3969/j.issn.1673-629X.2018.05.001

Research on Data Preprocessing Methods for Big Data

KONG Qin,YE Chang-qing,SUN Yun
(Nanjing University,Nanjing 210089,China)

Abstract: In the era of big data,it is an enormous challenge about data perception,expression,understanding and computing due to the inherent complexity of data type,organization pattern,different relations and data quality.Data preprocessing is a very important preparation before data analysis and mining.On the one hand,it ensures the correctness and effectiveness of data mining.On the other hand,the adjustment of the data format and content makes data meet the demand of mining.We analyze the main tasks of data preprocessing and summarize several popular processing methods for handling various kinds of “dirty data”.The algorithms of data cleaning,integration,transformation and reduction are discussed in detail.Using such kinds of preprocessing methods,we can remove redundant and error data,improve the incomplete data,promote the required data integration,help data refinement and data consistency of centralized storage.We also can get the minimum and the most reliable data set necessary for the mining system.It also reduces the cost of data mining and improves the accuracy,validity and practicability of knowledge discovery.

Key words: big data; preprocessing; dirty data; research

0 引 言

大数据中蕴含的宝贵价值成为人们存储和处理大数据的驱动力。在《大数据时代》一书中指出了大数据时代处理数据理念的三大转变,即要全体不要抽样,要效率不要绝对精确,要相关不要因果[1]。海量数据的处理对于当前存在的技术来说是一种极大的挑战。大数据的涌现使人们处理计算问题时获得了前所未有的大规模样本,但同时也不得不面对更加复杂的数据对象。数据预处理作为数据分析、挖掘前的重要数据准备工作,可以保证数据挖掘结果的准确性和有效性。

1 研究背景

大数据环境下,来自异构系统的原始数据中存在若干问题:

(1) 杂乱性。原始数据是从各个实际应用系统中获取的,由于各应用系统的数据缺乏统一标准的定义,数据结构也有较大的差异,因此各系统间的数据存在较大的不一致性,往往不能直接拿来使用。

(2) 重复性。是指对于同一个客观事物在数据库中存在其两个或两个以上完全相同的物理描述。这是应用系统实际使用过程中普遍存在的问题,几乎所有

收稿日期: 2017-04-13 **修回日期:** 2017-08-15 **网络出版时间:** 2018-02-07
基金项目: 国家自然科学基金(90412014);全国高等院校计算机基础教育研究会计算机基础教学研究与改革课题(AFCEC-2016-18)
作者简介: 孔 钦(1983-),女,讲师,硕士研究生,研究方向为计算机应用、数据分析。
网络出版地址: http://kns.cnki.net/kcms/detail/61.1450.TP.20180207.1525.010.html

应用系统中都存在数据的重复和信息的冗余现象[2]。

(3)模糊性。由于实际系统设计时存在的缺陷以及一些使用过程中的人为因素,数据记录中可能会出现有些数据属性的值丢失或不确定的情况,还可能缺失必需的数据而造成数据不完整。在实际使用的系统中,存在大量的模糊信息,有些数据甚至还具有一定的随机性质。

如前所述,因为数据类型和组织模式多样化、关联关系繁杂、质量良莠不齐等内在的复杂性,使得数据的感知、表达、理解和计算等多个环节面临着巨大的挑战。因此,数据预处理是数据挖掘前的一个非常重要的数据准备工作,是知识发现过程(knowledge discovery in database,KDD)的关键环节之一[3]。一方面它可以保证挖掘数据的正确性和有效性,另一方面通过对数据格式和内容的调整,使数据更符合挖掘的需要。通过把一些与数据分析、挖掘无关的数据项清除掉,为挖掘算法提供更高质量的数据内核。

数据挖掘的首要前提是确保消除所有的“脏数据”,包含冗余数据、缺失数据、不确定数据和不一致数据。针对“脏数据”的预处理方法有以下几种:清洗、集成、变换和归约。

1.1 数据清洗

检测数据中存在冗余、错误、不一致等噪声数据,利用各种清洗技术,形成“干净”的一致性数据集合。如图1所示。

图1 数据清洗

数据清洗技术包括清除重复数据、填充缺失数据、消除噪声数据等。在分析“脏数据”的产生来源和存在形式后,充分利用新兴的技术手段和方法去清洗“脏数据”,将“脏数据”转化为满足数据质量或应用要求的数据。美国最早对数据清洗技术展开研究。随着信息业和商业的发展,数据清洗技术得到了进一步发展。数据清洗分为以下几大类:

(1)重复数据的清洗。为了提高数据挖掘的速度和精度,有必要去除数据集合中的重复记录。如果有两个及以上的实例表示的是同一实体,那么即为重复记录。为了发现重复实例,通常的做法是将每一个实例都与其他实例进行对比,找出与之相同的实例。对于实例中的数值型属性,可以采用统计学的方法来检测,根据不同的数值型属性的均值和标准方差值,设置不同属性的置信区间来识别异常属性对应的记录,识别出数据集合中的重复记录,并加以消除。相似度计算是重复数据清洗过程中的常用方法,通过计算记录的各属性的相似度,再考虑每个属性的不同权重值,加权平均后得到记录的相似度。如果两条记录相似度超过了某一阈值,则认为两条记录是匹配的,否则,认为这两条记录指向不同实体[4]。另一种相似度计算算法基于基本近邻排序算法。核心思想是为了减少记录的比较次数,在按关键字排序后的数据集上移动一个大小固定的窗口,通过检测窗口内的记录来判定它们是否相似,从而确定重复记录。

(2)缺失数据清洗(missing values imputation)。完善缺失数据是数据清洗领域面临的另一个重要问题。如图2所示,在现实世界中,由于手动输入的失误操作、部分信息需要保密或者数据来源不可靠等各种各样的原因,使得数据集中的内容残缺不完整。比如某条记录的属性值被标记为NULL\、空缺或“未知”等。一旦不完整、不准确的数据用于挖掘,则会影响抽取模式的正确性和导出规则的准确性。当错误的数据挖掘模型应用于前端的决策系统时,就会导致分析结果和执行决策出现严重偏差[5]。

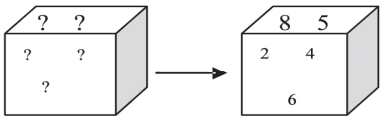

图2 缺失数据清洗

当前有很多方法用于缺失值清洗,可以分为两类:

(a)忽略不完整数据。直接通过删除属性或实例,忽略不完整的数据[6]。在数据集规模不大、不完整数据较少的情况下,常常利用该方法来实现数据清洗。该方法因为执行效率高,因此经常作为缺省方法,但缺点也相当明显。如果不完整数据集较大,一旦删除了若干记录之后,因为剩余的数据集规模较小,使得模型的构建不具备普适性和代表性,无法让人信赖,可靠度大大降低。另外,因为删除不完整数据带来的数据集偏差也使得数据挖掘的分类、聚类模型产生严重倾斜,进而影响最终的挖掘结果,产生重大决策性误导。

(b)基于填充技术的缺失值插补算法。上一种忽略法很有可能将潜在的有价值信息也一并删除。因此更多的时候选择填充不完整的数据。为了填充缺失值,用最接近缺失值的值来替代它,保证可挖掘数据的数量和质量。填充方法保留了潜在的有用数据,和删除属性或记录相比,保留了更多数据样本,不易于产生数据分析偏差,由此构建的模型更可靠,更有说服力。

目前常用的缺失值填充算法大体分为两大类,一类是统计学方法,另一类是分类、聚类方法。

·采用统计学方法填充缺失值。分析数据集,获取数据集的统计信息,利用数值信息填充缺失值。其中最简单的方法是平均值填充方法[7]。它把所有完整数据的算术平均值作为缺失数据的值。这种方法的弊

端在于有可能会影响缺失数据与其他数据之间原本的相关性。如果规模较大的数据集的缺失值全部采用平均值填充法进行填充,因为过多的中值存在,更多的尖峰态频率分布有可能会误导挖掘结果。

·采用分类、聚类方法填充缺失值。分类是在已有类标号的基础上,通过输入训练样本数据集,构造出分类器(如分类函数或者分类模型)。常用的数据分类技术包括决策树、神经网络、贝叶斯网络、粗糙集理论、最临近分类法等。利用完整记录与缺失记录之间的记录相似度,通过最大相似度的计算,结合机器学习的相关技术,建立最大可能的完整的数据模型。聚类是在不考虑类标号的前提下,寻求类间的相似性,目的也是在海量的数据聚集的基础上,构建较小的代表性的数据集,并基于该集合进一步分析和研究。常见的缺失值填充算法包括 EM 最大期望值算法(expectation-maximization algorithm)、MI 算法(multiple imputation)和 KNNI 算法(k-nearest neighbor imputation)等。其中最大期望算法通过创建概率模型,寻找参数最大似然估计值或者最大后验估计值,概率模型的成功与否依赖于无法观测的隐藏变量(latent variable)[8-9]。

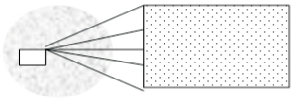

图 3　噪声数据

(3) 噪声数据处理(noise treatment)。数据挖掘前,往往假设数据集不存在任何数据干扰。然而,实际应用中却因为各种原因,在数据收集、整理的过程中,产生大量的噪声数据,即"离群点"。因为噪声数据不在合理的数据域内,所以分析、挖掘过程中输入和输出数据的质量难以保证,容易造成后续的挖掘结果不准确、不可靠,如图 3 所示。常用的消除噪声数据的方法分为两种。一种叫噪声平滑方法(data polishing),常用的方法是分箱法。将预处理数据分布到不同的箱中,通过参考周围实例平滑噪声数据,包括等宽分箱和等深分箱两大类。具体的分箱技术包括:按箱平均值平滑,即求取箱中的所有值的平均值,然后使用均值替代箱中所有数据;按中位数平滑,和上一种方法类似,采用中位数进行平滑;按设定的箱边界平滑,定义箱边界是箱中的最大和最小值。用最近的箱边界值替换每一个值。另一种是噪声过滤(data filters),利用聚类方法对离群点进行分析、过滤。在训练集中明确并去除噪声实例。噪声过滤的常用算法包括 IPF 算法(iterative partitioning filter)、EF 算法(ensemble filter)[10]。

1.2　数据集成

数据集成(data integration)是将多文件或多数据库运行环境中的异构数据进行合并处理,解决语义的模糊性。该部分主要涉及数据的选择、数据的冲突问题以及不一致数据的处理问题,如图 4 所示。

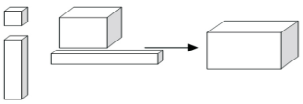

图 4　数据集成

1.3　数据变换

数据变换(data transformation):是找到数据的特征表示,用维变换或转换来减少有效变量的数目或找到数据的不变式,包括规格化、切换和投影等操作。数据变换是将数据转换成适合于各种挖掘模式的形式,根据其后所使用的数据挖掘算法,决定选择使用何种数据变换方法。常用变换方法包括:函数变换,使用数学函数对每个属性值进行映射;对数据进行规范化,按比例缩放数据的属性值,尽量落入较小的特定区间。规范化既有助于各类分类、聚类算法的实施,又避免了对度量单位的过度依赖,同时规避了权重不平衡发生。

1.4　数据归约

数据归约(data reduction):是在对发现任务和数据本身内容理解的基础上,寻找依赖于发现目标的表达数据的有用特征,以缩减数据模型,从而在尽可能保持数据原貌的前提下最大限度地精简数据量,促进大数据挖掘更高效。其主要有两个途径:维归约和数量归约,分别针对数据库中的属性和记录。目前海量数据上的数据归约技术是数据预处理的重要问题之一。

归约过程涉及的重要技术包括:

(1) 针对高维数据的降维处理(dimensionality reduction)。涉及的技术包括特征值选择(feature selection)和空间变换(space transformations)。维归约的核心是减少随机变量或者属性的个数。特征值选择目的是获取能描述问题的关键特征的那部分属性。删除不相关的、冗余的属性,使得机器学习过程更快,内存消耗更少。特征子集选择方法,包括各类启发式算法、贪心算法等,具体有向前选择法、向后删除法、决策树归纳法等。数量归约的重点在于减少数据量,从数据集中选择较小的数据表示形式。主流的数值归约技术,包括对数线性模型、直方图、聚类、抽样等。常用算法包括:LVF(Las Vegas filter)、MIFS(mutual information feature selection)、mRMR(minimum redundancy maximum relevance)、Relief 算法。空间变化是另一种降低数据维度的方法。流行的算法有 LLE(locally linear embedding)、PCA(principal components analysis)等[11]。

(2) 实例归约(instance reduction)。当前很流行的一种减少数据集规模的算法是实例归约算法。在减少数据量的同时,并没有降低获取知识的品质。通过移除或者生成新的实例的方法,大大降低了数据规模。

涉及技术包括:(a)实例选择(instance selection)。好的实例选择算法能够生成一个最小的数据集,移除噪声数据和冗余数据,独立于随后进行的数据挖掘算法,符合数据分析和挖掘的要求。常见的算法有 CNN(condensed nearest neighbor)、ENN(edited nearest neighbor)、ICF(iterative case filtering)、DROP(decremental reduction by ordered projections)等。(b)实例生成(instance generation)。建立各种原型用于实例生成,涉及算法包括 LVQ(learning vector quantization)等[12]。

(3)离散化技术(discretization)。目的在于减少给定连续属性值的个数。离散化之前,首先要预估离散型数据的规模,接着对连续型数据进行排序,然后指定若干个分裂点把数据分为多个区间。将落在同一个区间内的所有连续型数据通过统一的映射方法对应到相同的离散型数据上[13]。根据分裂点认定方式的不同,离散化分为自顶向下和自底向上两种,按照是否使用分类信息,又分为监督和非监督两大类。目前大多数离散化方法分为两大方向,一是从属性出发,基于属性的重要性进行离散处理,二是利用分辨矩阵进行映射。常见的算法包括:MDLP(minimum description length principle)、ChiMerge、CAIM(class-attribute interdependence maximization)等[14]。

(4)不平衡学习(imbalanced learning)。在使用机器学习的有监督学习形成数据模型时,很容易在不同类型的数据集上产生巨大的优先级的差异。这种也叫做分类不平衡问题。很多标准的分类学习算法经常会倾向于大多数实例(majority class)而忽视少数特别实例(minority class)[15]。数据预处理相关技术可以避免出现类型分布不平衡的情况。主要方法是两种:欠采样方法,在抽样创建原始数据集的子集用作数据挖掘时,尽量去除大多数实例;过度采样方法,在抽样时复制很多相同的实例或者创建新的实例。在众多采样算法中,最复杂最著名的遗传算法是 SMOTE(synthetic minority oversampling technique)。

2 结束语

大数据时代下,不同的应用领域、各种新兴的云计算技术会促进数据预处理方法进一步的扩展和提升。数据预处理是知识发现过程中十分重要的环节,是数据挖掘算法能够有效执行的必要前提。通过高效的预处理工作,清除冗余数据,纠正错误数据,完善残缺数据,挑选出必需的数据进行集成,达到数据信息精练化、数据格式一致化和数据存储集中化。在最精确、最可靠的数据集合上进行数据挖掘,极大地减少了系统挖掘的开销,提高了知识发现的准确性、有效性和实用性。

参考文献:

[1] 程学旗,靳小龙,王元卓,等.大数据系统和分析技术综述[J].软件学报,2014,25(9):1889-1908.

[2] 李小菲.数据预处理算法的研究与应用[D].成都:西南交通大学,2006.

[3] WU X,ZHU X,WU G Q,et al.Data mining with big data[J].IEEE Transactions on Knowledge and Data Engineering,2014,26(1):97-107.

[4] GARCÍA S,LUENGO J,HERRERA F.Tutorial on practical tips of the most influential data preprocessing algorithms in data mining[J].Knowledge-Based Systems,2016,98:1-29.

[5] 关大伟.数据挖掘中的数据预处理[D].长春:吉林大学,2006.

[6] TRIGUERO I,PERALTA D,BACARDIT J,et al.MRPR:a MapReduce solution for prototype reduction in big data classification[J].Neurocomputing,2015,150:331-345.

[7] GALAR M,FERNÁNDEZ A,BARRENECHEA E,et al.A review on ensembles for the class imbalance problem:bagging-,boosting-,and hybrid-based approaches[J].IEEE Transactions on Systems,Man,and Cybernetics,Part C:Applications and Reviews,2012,42(4):463-484.

[8] GAO M,HONG X,CHEN S,et al.A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems[J].Neurocomputing,2011,74(17):3456-3466.

[9] SOTOCA J M,PLA F.Supervised feature selection by clustering using conditional mutual information-based distances[J].Pattern Recognition,2010,43(6):2068-2081.

[10] MITRA P,MURTHY C A,PAL S K.Density-based multiscale data condensation[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2002,24(6):734-747.

[11] WANG H,WANG S.Mining incomplete survey data through classification[J].Knowledge and Information Systerms,2010,24(2):221-233.

[12] PÉREZORTIZ M,GUTIÉRREZ P A,MARTÍNEZ C H,et al.Graph-based approaches for over-sampling in the context of ordinal regression[J].IEEE Transactions on Knowledge and Data Engineering,2015,27(5):1233-1245.

[13] PRATI R C,BATISTA G E A P A,SILVA D F.Class imbalance revisited:a new experimental setup to assess the performance of treatment methods[J].Knowledge and Information Systems,2015,45(1):247-270.

[14] ANGIULLI F,FOLINO G.Distributed nearest neighbor-based condensation of very large data sets[J].IEEE Transcactions on Knowledge and Data Engineering,2007,19(12):1593-1606.

[15] BACARDIT J,WIDERA P,CHAMORRO A E M,et al.Contact map prediction using a large-scale ensemble of rule sets and the fusion of multiple predicted structural features[J].Bioinformatics,2012,28(19):2441-2448.