基于 KM-SMOTE 和随机森林的不平衡数据分类

陈 斌,苏一丹,黄 山
(广西大学 计算机与电子信息学院,广西 南宁 530004)

摘 要:基于 SMOTE 算法的随机森林能够很好地处理不平衡数据集的分类,是一种通过对数据进行改造以达到良好分类要求的分类器。但 SMOTE 算法在处理不平衡数据后,可能会导致不平衡数据集分布的整体变化以及模糊正负类边界。这两个缺陷极易导致平衡后的数据与原始数据集有很大差异,从而使分类结果有提高但仍旧不够理想。K-means 算法能够有效地聚类,并达到对数据分布的描述。在此基础上,结合 K-means 算法与 SMOTE 算法,利用两者优点,文中提出了一种基于 K-means 的 KM-SMOTE 算法,有效地解决了上述两个问题。并用于随机森林分类器进行实验,结果表明,改进后的算法分类效果更加明显。

关键词: K-means;SMOTE 算法;随机森林;不平衡数据集
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2015)09-0017-05
doi:10.3969/j.issn.1673-629X.2015.09.004

Classification of Imbalance Data Based on KM-SMOTE Algorithm and Random Forest

CHEN Bin, SU Yi-dan, HUANG Shan
(College of Computer Science and Electronics Information, Guangxi University, Nanning 530004, China)

Abstract: The random forest based on SMOTE algorithm can be a good deal classification in imbalance data, is a classifier through transforming the data to achieve good classification requirements. But after SMOTE algorithm deals with imbalance data, may cause overall changes of the distribution of imbalance data sets, and fuzzy the boundaries of positive class and negative class. Both defects can easily lead to big difference from the balanced data sets and the original data sets after the change, resulting in classification results not satisfactory. The K-means clustering algorithm can effectively cluster and describe the data distribution. On this basis, combined with K-means algorithm and SMOTE algorithm, using the advantages of both, present a KM-SMOTE algorithm based on K-means algorithm, successfully resolving these two issues. And for random forest classifier make an experiment. The results also demonstrate that the effect of the improved classification algorithm is more obvious.

Key words: K-means; SMOTE algorithm; random forest; imbalance data set

0 引 言

分类是根据数据集的特点构造一个分类器,利用分类器对未知类别的样本赋予类别的一种技术[1]。分类是数据挖掘中重要内容之一,它包括两个阶段:学习阶段和分类阶段[2]。学习阶段的内容为建立描述预先定义的数据类或概念集的分类器,通过对训练集的“学习”来构造分类器。而在分类阶段,以构造完毕的分类器为标准,将未分类数据依照分类器进行数据分类操作,得出最后的分类结果。对于数据的分离方法,根据输入区间的划分大概可分为以下三大类:聚类划分、决策树划分以及网格划分[3]。而现阶段比较成熟且较为流行的分类算法包括以下几种:贝叶斯、决策树、支持向量机、神经网络以及 KNN 等。其中,决策树分类算法是从 J. R. Quinlan 于 1986 年发文介绍 ID3 算法开始开发而完善起来,经历了如 C4.5、CART、CHAID 等相应阶段,现已有多种不同算法模型。但不论哪一种决策树模型,都离不开同样的分类原理,以相应的方法选取最优属性,进行节点分裂。而随机森林(Random Forests)便是基于决策树算法的集成学习分类模型。

收稿日期:2014-11-04 **修回日期**:2015-02-06 **网络出版时间**:2015-08-26
基金项目:教育部人文社会科学研究项目(11YJAZH080)
作者简介:陈 斌(1988-),男,硕士研究生,研究方向为数据挖掘、数据库理论;苏一丹,教授,博士,研究方向为数据挖掘、电子商务、信息安全。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20150826.1558.056.html

随机森林是由美国科学家 Leo Breiman 于2001 发表的一种机器学习方法[4],其提出原理是结合 Bagging 集成学习理论和随机子空间方法。因随机森林是一个集成学习模型,并以决策树为基本分类器,故随机森林包含多个决策树,通过多个决策树对输入样本的分类判定,投票决定最终的分类结果。随机森林算法能够克服决策树的过拟合,能够较好地容忍异常值和噪声,还有处理高维数据时的并行性和可扩展性。学者们在 Leo Breiman 的基础上又对随机森林做了许多改进,如 Hemant Ishwaran 等提出的随机生存森林[5]。同样,也有用随机森林处理不平衡数据集的研究,如房晓南等针对 Web spam 的检测[6]。

对不平衡数据集的分类,因为其数据稀少、数据碎片、归纳偏差以及噪声等问题,处理起来比较困难。针对这些缺陷,通过改进算法以及改造数据能够达到较好的效果。SMOTE 算法是 2002 年 Chawla 等提出来的[7],该算法的本质是对随机向上抽样策略的改进。很多人利用这种算法来处理不平衡数据集,如 Han 等提出的 Borderline-SMOTE 算法[8],或曾志强等提出的基于核 SMOTE 的 SVM 来处理不平衡数据[9]等。但由于 SMOTE 算法有其自身的缺陷,除了无法更好地体现非平衡数据的分布之外,还有严重的数据重复以及数据修正之后的正负类边界模糊问题。文中正是在此基础上,提出了能够解决问题的 KM-SMOTE 算法,结合随机森林在处理不平衡数据集时达到更好的分类效果。

1 相关知识

1.1 *K*-means 算法

在数据挖掘中,*K* - means 主要是用来计算数据聚集,通过不断计算离初始簇心近的点来聚类的算法。

对于一组测试数据集 $\{x_1,x_2,\cdots,x_n\}$,每个测试数据是一个 h 维向量,从数据集中以特定的方法取 K 个数据作为起始簇心,每一个数据点代表一个簇,则共有 K 个簇 $\{c_1,c_2,\cdots,c_k\}$。其他任意测试数据 $p \in c_i(i=1,2,\cdots,k)$ 与该簇的关系程度用欧氏距离来表示,使分类过程满足 $\min\limits_{c}\sum\limits_{i=1}^{k}\sum\limits_{x_j\in c_i}\| x_j - u_i \|^2$,其中 u_i 是 c_i 的簇心。

得到完整的 K 个簇后,重新计算簇心,并以此替代原有的簇心,重复上述过程,直到达到最大迭代次数或者两个欧氏距离之间的差别小于一个给定阈值。

该算法有两个点值得关注,一是 K 值的选取,二是随机簇心的落处。对于一个数据集,由于事先并不知道具体的合适聚类数量,K 值非常难以估计,这会影响到算法本身的准确性。而簇心是随机选取,不同的随机簇心点也会对结果有影响。

1.2 SMOTE 算法

不平衡数据,即数据集中某一类的样本数量明显少于其他类的样本数量,根据数量的多少,分别对应多数类和少数类。对不平衡数据集样本,由于分类数据集的特点,决定了其在分类过程中难以满足分类的准确性要求,由于极端值、噪声和数据稀缺等一系列问题,导致数据分类方法的性能出现极大的下降,甚至达到不可接受的地步。

SMOTE 算法通过改变数据集的平衡性来维持数据分类方法的性能,即利用增加少数类数据的方法,使其与多数类的数据样本达到一个平衡。这种方式效果极佳,还有以核学习为基础,SMOTE 算法为主线,针对复杂度研究提出整体解决方案,提高算法在应用方面的需求[10]。

SMOTE 算法,针对少数类数据中的每一个数据样本 X,搜索它的最近邻样本 K 个,再在这些最近邻数据集中随机选择 N 个样本。则对每一个原始样本数据,选择了 K 个近邻样本中的 N 个样本,接下来便是在原始样本数据以及它的近邻样本之间进行插值操作。公式如下:

$$X_{new} = X + \text{rand}(0,1) * (M_i - X), i = 1,2,\cdots,N \tag{1}$$

其中,X_{new} 为新插值的样本;X 为选择的原始样本数据;$\text{rand}(0,1)$ 表示 0 与 1 之间的的某一随机数;M_i 为原始样本数据 X 的最近邻 K 个样本中选取的 N 个样本[7]。

1.3 随机森林

随机森林是用来解决预测问题的学习模型[11]。以 K 个决策树为基础利用集成学习模型生成随机森林,每一棵决策树为一个基本分类器,随机森林的输出结果依赖于每一棵决策树的分类结果,采用简单的投票方式来决定。

K 个决策树的描述为:

$$\{h(X,\theta_k), k = 1,2,\cdots,K\} \tag{2}$$

其中,K 为随机森林所包含的决策树个数;$\{\theta_k\}$ 代表独立同分布的随机向量。

针对自变量 X,K 个决策树都会进行分类,然后选出最优分类结果。

分类结果描述为:

$$H(x) = \arg\max_{Y}\sum_{i=1}^{k} I(h_i(x) = Y) \tag{3}$$

其中,$H(x)$ 表示随机森林的组合分类模型,会给出自变量 X 的最终分类结果;$I(*)$ 为示性函数,用来表示决策树的分类;h_i 表示单个分类模型;Y 是决策树分类结果的输出变量。

对于随机森林模型,因其是多个决策树的集成,故它的训练和分类过程可以看成是多个决策树训练和分类的集合。在训练和分类过程中,由于每一棵决策树的训练和分类过程是独立的,故可以通过并行处理来降低程序运行的时间。

2　基于 KM-SMOTE 算法的随机森林

2.1　算法的提出

对于不平衡数据集的分类问题,上文提到了现今非常常用的 SMOTE 算法,对此算法,分析它的插值公式,便可得出如下结论:

结论 1:SMOTE 算法在插值过程中,从所有稀有原始数据中随机选择,概括来说,SMOTE 算法对稀有数据类的所有原始数据样本一视同仁,随机进行插值操作。这个过程可能导致插入数据之后数据分布原型的改变,而不满足最初的数据集分布模型。

结论 2:假设新插值对象的原始数据样本处于少数类样本边界,根据插值公式,若 rand(0,1) 取值使得 X_{new} 的值正向变大,极有可能导致新插值对象进一步向多数类样本靠拢,而模糊正负类样本边界。

算法的这两处缺陷都有可能导致数据集进行平衡操作后而改变数据本身的分布状态,出现大的偏差,也导致分类结果不够理想。

所以针对这两点缺陷,很多研究人员做了改动,如曹正凤提出的 C-SMOTE 算法[12],利用样本中心调整插值结果。文中正是在各项研究基础上,根据 K - means 算法以及 SMOTE 算法的本质,提出了基于 K - means 的 KM-SMOTE 算法,并结合随机森林来处理不平衡数据集进行分类。

2.2　算法原理

对于 SMOTE 算法的两个不足的地方,即插入数据过程中的一视同仁以及模糊多数类和少数类边界,若要完善此两项缺陷,需要考虑在插值过程中,尽可能地以区域分布为特征插入数据项,以及新插入的数据值,不能存在模糊原始数据集最边界的数据的插值。

KM-SMOTE 算法在进行插值前,一是先将数据集进行聚类操作,这样以聚类为区域进行插值,能够有效地防止插值泛化,而出现改变数据集分布特征的情况;二是 KM-SMOTE 算法的插值公式经过修正,所插值的数据在簇心和原始数据点的连线上,不会出现泛边界的数据。

KM-SMOTE 算法的原理如下:对少数类数据集,首先利用 K 均值算法进行聚类,这样有助于数据集形成以簇为中心的数据聚集,帮助有针对性地进行插值。聚类结束后,会形成固定的 K 个簇并记录每一个簇心,以簇心为插值的原始样本点,针对每个簇的样本进行插值操作。

2.3　算法设计

(1)初始簇 K 值的确定。在以往 K - means 算法中,K 值的大小一般是程序需要指定的某个值或者利用经验来设定一个值,这样取的 K 值简单有效。

(2)利用 K - means 算法聚类并计算簇心。选择稀有数据类样本,利用 K - means 算法进行聚类操作并记录簇心。稀有数据类共分为 K 个聚类,每一个聚类的簇心为 $\{c_1,c_2,\cdots,c_k\}$。

(3)进行样本插值。为了有效降低数据的偏向性以及改变数据的平衡性,文中是利用簇心而非稀有数据集内的原始样本点进行插值,即依照文中所提出的新算法—KM-SMOTE 算法。新插值公式设定如下:

$$X_{new} = c_i + \text{rand}(0,1) * (X - c_i), i = 1,2,\cdots,k, \quad X \in c_i \tag{4}$$

其中,X_{new} 为新插值的样本;c_i 为簇心;X 是以 c_i 为簇心聚类中的原始样本数据;rand(0,1) 表示 0 与 1 之间的某一随机数。

(4)处理插值后的数据集。对每一个簇心等概率的样本插值后,稀有类数据集中的样本可能比非稀有类数据集中的样本要多,此时要进行样本删除操作。数据删除的方法是删除每一个簇中可能产生过拟合的数据,直到达到数据集平衡。

(5)使用随机森林进行分类。处理过平衡后的数据集之后,利用该平衡数据集进行训练和分类。进行分类结果的记录和分析,即可得到最新算法的实验效果。

因在实验中,要进行实验的对比,故在算法设计过程,以原始数据直接进行分类的数据可以直接进行第五步操作。而以 SMOTE 算法处理的数据,则需要将前三步置换成 SMOTE 算法处理数据的过程。

3　实　验

文中实验所采用的数据集为源于 UCI 机器学习数据库的 Yeast、Pima 以及 Seg 数据集,数据可以从 UCI 数据库中下载得到,该三项数据集为明显不平衡数据集。

数据集如表 1 所示。

表 1　初始不平衡数据集列表

数据集	样本数	变量数	少类样本数	多类样本数	不平衡比/%
Yeast	1 484	9	51	1 433	3.56
Pima	768	8	268	500	53.5
Seg	2 310	19	330	1 980	16.7

由表中可以看出,每一个数据集的样本数分为两类,即多数类样本和少数类样本。不平衡比是少数类

样本数与多数类样本数之间的比值,比值越小说明数据集不平衡性越大。

将此三项数据集按照文献[13]的方法以 6:1 的比率随机分为训练集和测试集,并运行 100 次取平均值来进行结果判定。

3.1 实验步骤

对于实验,根据算法的具体情况进行针对实施,首先要对采用的数据集进行数据预处理,架构实验环境,编写算法代码,进行数据分类操作。具体实施的步骤如下所示:

(1)确定数据集。对 Yeast、Pima、Seg 进行预处理操作,按照实验要求设定实验中的各种参数。

(2)进行 K - means 算法的聚类操作。在 K - means 算法中,影响结果最重要的两个原因都跟随机性有关,即 K 值的选取以及初始簇心的随机分配。此实验,根据非平衡数据集的规模,随机选取 5 个点作为初始簇心进行聚类操作,直到达到最大迭代次数或者两个欧氏距离之间的差别小于一个给定阈值,算法结束,记录最终簇心点。

(3)对以上簇心,进行 KM-SMOTE 算法的插值操作。出于算法改正的考虑,为了验证实验结果,需要对使用了 K - means 算法后的数据集用新算法 KM-SMOTE 进行插值操作,操作依据公式(4)。根据新的插值公式插值的数据是为了降低原 SMOTE 算法所存在的缺陷。数据集平衡后稍作处理,以使正负类数据集达到平衡。

(4)利用随机森林对平衡后的数据集进行分类操作,对插值之后的数据用随机森林进行分类。在这里,需要提到的是,分类包括三个部分,分别是:原数据集分类,经 SMOTE 算法处理过的数据集分类以及经 KM-SMOTE 算法处理后的数据集分类。记录最终结果并做出总结与评价。

3.2 评价指标

评价指标采用分类器性能判断的常用指标 G-means 准则[13]。作为不平衡数据集学习中的评价标准,G-means 的取值取决于多数类精确度和少数类精确度乘积的平方根。只有当两者的值都较大时,即多数类与少数类的分类精确度都较高时,G-means 的取值才会大。可以说,几何均值 G-means 是能够合理评价不平衡数据集总体分类性能的一个标准。

G-means 判断准则的描述如下,所有测试数据集最后分为四种情况,正确分类的多数类、正确分类的少数类以及错误分类的多数类、错误分类的少数类,分别表示为:TN、TP、FP、FN,见表 2。

因 G-means 准则的评价方法取决于两个度量,分类器对少数类分类的准确性以及分类器对多数类分类的准确性。可以简约地用两个值 a^+ 与 a^- 来表述。

表 2 G-means 准则矩阵

	判断为少类	判断为多类
真实为少类	TP	FN
真实为多类	FP	TN

表示少数类样本分类的准确率:

$$a^+ = TP/(TP + FN) \tag{5}$$

表示多数类样本的分类准确率:

$$a^- = TN/(TN + FP) \tag{6}$$

总体分类性能指标:

$$G - means = \sqrt{a^+ \times a^-} \tag{7}$$

通过计算 G-means 的最终数值,来评价分类器的分类效果,数值越大的分类效果越好。

3.3 实验结果及分析

实验中随机选取训练集以及测试集后,对训练集利用 KM-SMOTE 进行数据插值操作以达到数据平衡,设计随机森林分类器,进行分类操作,并将分类结果进行对比。

以横坐标表示不同处理方式的数据集进行分类操作,纵坐标表示不同处理方式的数据集分类之后的 G-means 数值。

将三者分类结果进行对比,Yeast、Pima、Seg 的对比结果分别如图 1 ~3 所示。

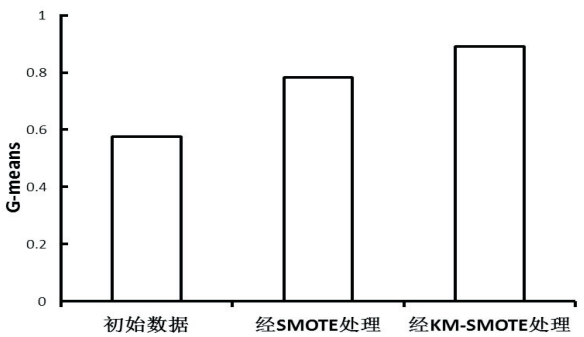

图 1 Yeast 数据集的对比结果

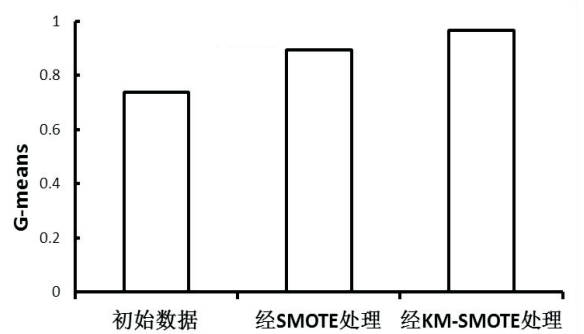

图 2 Pima 数据集的对比结果

对比内容除了使用 KM-SMOTE 算法处理后的不平衡数据的分类,还包括另外两方面:一是未经过数据处理的数据集,用随机森林方法直接进行分类;二是用

基于常规 SMOTE 算法对不平衡数据集进行数据平衡操作,继而用随机森林法进行分类。

从图 1 中可以看出,对于数据集 Yeast,未经处理直接用随机森林法进行分类,分类结果的 G-means 值为 0.574 8;经过 SMOTE 算法平衡数据集后,分类结果的 G-means 值为 0.782 3;而经过 KM-SMOTE 算法平衡数据集后,分类结果的 G-means 数值为 0.891 5。忽略掉 K - means 算法本身随机性所带来的影响,依旧可以看出,KM-SMOTE 算法的分类效果比其他两种算法的分类效果要好。

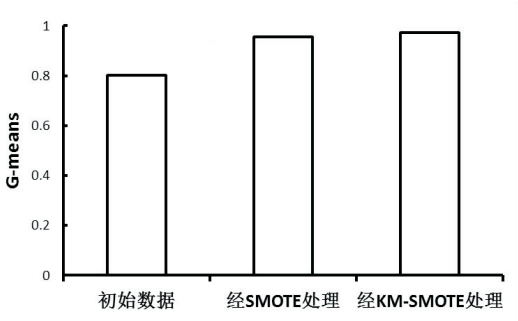

图 3　Seg 数据集的对比结果

从图 2 和图 3 也可以看出,除了 Yeast 数据集分类效果可以证明使用 KM-SMOTE 算法能得到更好的分类结果外,Pima 和 Seg 数据集的分类结果同样可以说明此项结论。

结合三个数据集最后的分类效果进行类比,结果如图 4 所示。

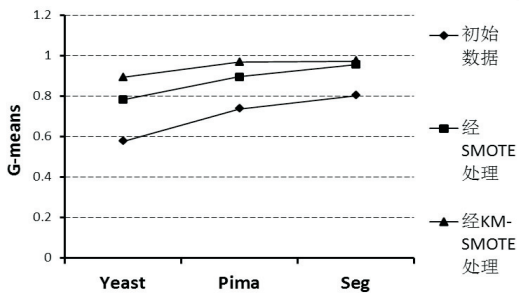

图 4　三组数据的对比结果

可见,基于 KM-SMOTE 的随机森林在不平衡数据集上的分类结果相比其他两种方法有很大的优势,这表示文中提出的算法改进是有效的。

4　结束语

SMOTE 算法作为处理不平衡数据集最经典的算法,由于其在改变数据集方面的性能表现,一直被学术界加以改进与研究。但部分研究的方向是让 SMOTE 算法进一步提高平衡数据的效率和准确率, 很少针对 SMOTE 算法本身的缺陷来进行改进。文中基于 SMOTE 算法的随机森林分类器处理非平衡数据集的缺陷,与 K - means 算法相结合,提出了一种新的解决方案:基于 KM-SMOTE 算法的随机森林。该算法的随机森林分类器能够有效避免插值数据后导致的稀有类数据分布的变化,也解决了 SMOTE 算后模糊正负类边界的问题,并且实验也很好地证明了改进算法的有效性。但该算法依旧存在一些细小的缺陷,比如 K 值的选取和修正,运行速度的提高等。若在接下来的研究中,对该算法进行进一步改进,将会得到更好的结果。

参考文献:

[1] 刘红岩,陈　剑,陈国青. 数据挖掘中的数据分类算法综述[J]. 清华大学学报:自然科学版,2002,42(6):727-730.

[2] Han Jiawei, Kamber M, Pei Jian. 数据挖掘概念与技术[M]. 第 3 版. 北京:机械工业出版社,2012.

[3] 陈铁明,马继霞,Samuel H. Huang,等. 一种新的快速特征选择和数据分类方法[J]. 计算机研究与发展,2012,49(4):735-745.

[4] Breiman L. Random forests[J]. Machine Learning,2001,45(1):5-32.

[5] Ishwaran H, Kogalur U B, Blackstone E H, et al. Random survival forests[J]. The Annals of Applied Statistics, 2008, 2(3):841-860.

[6] 房晓南,张化祥,高　爽. 基于 SMOTE 和随机森林的 Web spam 检测[J]. 山东大学学报:工学版,2013,43(1):22-27.

[7] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16:321-357.

[8] Han Hui, Wang Wenyuan, Mao Binghuan. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[J]. Advances in Intelligent Computing, 2005, 3644: 878-887.

[9] 曾志强,吴　群,廖备水,等. 一种基于核 SMOTE 的非平衡数据集分类方法[J]. 电子学报,2009,37(11):2489-2495.

[10] 吴克寿,曾志强. 非平衡数据集分类研究[J]. 计算机技术与发展,2011,21(9):39-42.

[11] Verikas A, Gelzinis A, Bacauskiene M. Mining data with random forests: a survey and results of new tests[J]. Pattern Recognition, 2011, 44(2):330-349.

[12] 曹正凤. 随机森林算法优化研究[D]. 北京:首都经济贸易大学,2014.

[13] Wu G, Chang E. Class-boundary alignment for imbalanced data set learning[C]//Proc of workshop on learning from imbalanced data sets II. Washington DC:[s. n.],2003:49-56.