文章编号:1672 - 058X(2012)08 - 0036 - 06

基于 K-均值聚类的朴素贝叶斯网络分类模型

刘亚辉,王 越,谭暑秋
(重庆理工大学 计算机科学与工程学院,重庆 400054)

摘 要:针对朴素贝叶斯网络分类模型在处理高维大数据量时的效率偏低和准确率有待提高的问题,结合主元分析法与 K-均值聚类算法构造出了一个改进的朴素贝叶斯网络分类模型;摒弃了非类属性变量相对于类属性变量相对独立的前提条件,算法首先用主元分析法在对数据集的信息量尽量保存的同时进行了降维操作,使得算法可以着重于进行分类问题;算法还提出了一个"相对融合点"的概念,有效地提高了算法的性能;最后对算法的性能进行了分析,并将改进的算法应用到实际的数据集进行实验,用算法产生的分类结果对数据集中产生的一些缺失数据进行修补。

关键词:贝叶斯网络分类;朴素贝叶斯网络;K-均值聚类;数据挖掘
中图分类号:TP311.13 **文献标志码**:A

朴素贝叶斯网分类器(Naïve Bayesian classifier,NB)[1]是目前被公认的一种简单而有效的概率分类的方法,它同时也是贝叶斯网分类器的一种。从某种意义上来说其预测性能可与神经网络、决策树等算法相媲美,因此在某些领域中表现出了它性能的优异性[2]。许多工作研究者从不同的方向思考,并从 NB 方法中的"独立性假设"出发,为了提高分类器的性能从而构造出了不同的改进模型[3-5]。由于朴素贝叶斯网络的分类模型以各个非类属性变量相对于类属性变量相对独立为前提的,也就是各个非类属性变量独立地作用于类属性变量。此前提条件在一定程度上限制了朴素贝叶斯网络分类模型的适用范围,虽然降低了贝叶斯网络的构建复杂性,但是当处理的数据维数较多,且数据量较大时,朴素贝叶斯网络分类的效率则是偏低的,其准确率有待提高。在朴素贝叶斯网络的基础上结合主元分析法与 K-均值聚类算法构造出了一个改进的朴素贝叶斯网络的分类模型。由于主元分析法是解决多维数据行之有效的方法,而 K-均值聚类算法能够使簇内具有较高的相似度,而簇间的相似度较低,这将有便于从多维属性中有效地进行降维处理。

1 相关概念

概念 1(相对融合点) 在一个数据集 $D=(x_1,x_2,\cdots,x_n)$ 中,设存在一个值 x' 能够代表数据中的特点并与数据集具有很好的拟合性,则 x' 称为相对融合点。

概念 2(K-均值聚类) K-means 算法以 K 为参数,把 N 个对象分为 K 个簇,以使簇内的相似度较高,而簇间的相似度较低。根据一个簇中的平均值(视为簇重心)来进行相似度的计算。K-means 算法的处理过程如下:(1)随机选择 K 个对象,每一个对象初始代表一个簇的中心或平均值。计算剩余的每个对象与各个簇中心的距离,再

收稿日期:2012 - 02 - 12;**修回日期**:2012 - 03 - 15.
* **基金项目**:重庆市科技攻关资金资助项目(CSTC,2009AC2068).
作者简介:刘亚辉(1984-),男,河南驻马店人,硕士,从事数据库与数据挖掘研究.

将剩余的对象赋给最近的簇。(2) 不断重复地计算每个簇的平均值,直至准则函数收敛到期望值[6]。

概念 3(主元分析) 主元分析(PCA)是在有一定相关性的 m 个样本值与 n 个参数所构成的数据阵列的基础上,通过建立较小数目的综合变量,使其更集中的反映原有数据阵列中包含的信息的方法[7]。

2 算法描述

设存在一张数据表,有 m 个对象记为 $X_1, X_2, \cdots, X_m$,有 n 个属性记为 $a_1, a_2, \cdots, a_n$。其中每个对象 $X_i = (x_{i1}, x_{i2}, \cdots, x_{in})$, $i = 1, \cdots, m$,且 $x_{ij}(j = 1, \cdots, n)$代表对象 X_i 的 a_j 属性。同样 $a_j = (x_{1j,} x_{2j}, \cdots, x_{mj})$, $j = 1, \cdots, n$。因此这张表可以用下面的矩阵来表示:

$$B = [x_{ij}]_{m\times n} = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{pmatrix}$$

在一个历史数据集中总是会出现一些数据保存不完善的情况,而且丢失数据不只一个。为了使问题更容易地呈现,只讨论一个数据丢失数据的情况。当然此过程也可以被用到多个数据的丢失。改进的朴素贝叶斯算法的步骤如下:

假设 X_j 是缺损比较严重的数据列,属性 X_j 的值需要修补完整。在此假设数据的属性列比较多,算法的基本步骤如下:

第 1 步:利用 PCA 算法降维。

(1) 删除矩阵 B 的第 j 列,得到如下矩阵 D:

$$D = \begin{bmatrix} x_{11} & \cdots & x_{1j-1} & x_{1j+1} & \cdots & x_{1n} \\ x_{21} & \cdots & x_{2j-1} & x_{2j+1} & \cdots & x_{2n} \\ \vdots & & \vdots & \vdots & & \vdots \\ x_{m1} & \cdots & x_{mj-1} & x_{mj+1} & \cdots & x_{mn} \end{bmatrix}_{m\times(n-1)}$$

(2) 对样本数据进行标准化处理。

(3) 计算相关矩阵。对给定的 m 个样本,计算指标变量的相关系数矩阵:

$$R = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{pmatrix} = \frac{1}{m}X'X$$

其中 $r_{ij} = \frac{1}{m}\sum_{i=1}^{n} X_{ij}X_{ik} = \frac{1}{m}x'_j x_k$ 且 $j, k = 1, 2, \cdots, n$。

(4) 求特征值和特征向量。求解特征方程:$|R - \lambda f| = 0$。通过此方程,可得到 k 个特征值($i = 1 \sim n$)与对应于每一个特征值的特征向量 $Q_i = (a_{i1}, a_{i2}, \cdots, a_{ip})$,其中 $i = 1 \sim k$。

(5) 求主成分。通过上述方法可求得 $k(k \leqslant m)$ 个主成分。称 $\frac{\lambda_i}{\sum_{i=1}^{k}\lambda_i}$ 为第 i 个主成分的贡献率记为 β_i,且 $\beta_i = \frac{\beta_i}{\sum_{i=1}^{k}\lambda_i}$。在 k 个主成分中,称前 q 个主成分的贡献率之和为前 q 个主成分的累积贡献率,记为 α 且 $\alpha = \frac{\sum_{i=1}^{q}\lambda_i}{\sum_{i=1}^{k}\lambda_i}$。主成分的个数可以通过积累贡献率来确定,通常以累积贡献率 $\alpha \geqslant 0.85$ 为标准。

(6) 设通过 PCA 算法降维后的矩阵:

$$D_1 = \begin{bmatrix} x_{11} & \cdots & x_{1j-1} & x_{1j+1} & \cdots & x_{1n} \\ x_{21} & \cdots & x_{2j-1} & x_{2j+1} & \cdots & x_{2n} \\ \vdots & & \vdots & \vdots & & \vdots \\ x_{m1} & \cdots & x_{mj-1} & x_{mj+1} & \cdots & x_{mn} \end{bmatrix}_{m\times(n-l-1)}$$

第 2 步:求降维后各列属性值的相对融合点。

设列属性 $y_i = \{x_{1j}, x_{2j}, \cdots, x_{mj}\}$ 且 $j = 1, \cdots, n-l-1$,则 $D_1 = \{y_1, y_2, \cdots, y_n\}$。设列属性 y_i 中任何两个值之间的距离为 $d(x_{hi}, x_{ki})$。

(1) 求数值点的密度。确定一个正数 d_0,以每个数据为中心 d_0 为半径作 n 维空间的超球,令 $d(x_{hi}, x_{ki})$ 为 x_{hi}到 x_{ki}的距离,其中 $d(x_{hi}, x_{ki}) = \sqrt{|x_{hi} - x_{ki}|^2}(h, k = 1, 2, \cdots, m; j \neq k)$,若满足 $d(x_{hi}, x_{ki}) \leqslant d_0$ 则认为 x_{ki} 落在以 x_{hi}为圆心的超球内,落在超球内的点的总数为 x_{hi}的密度 p_{hi}。不难发现,p_{hi}越大,则以它为相对融合点的资格就越大。

(2) 计算列属性的相对融合点。在每列的各个球中,交集集合将归并到落在超球内的点的总数多的超球内。此时将落在各个超球内点的总数统计并按非递增排列,设为 $d_1, d_2, \cdots, d_m$,并相应地统计这些球内数据的总和,设为 $s_1, s_2, \cdots, s_m$。则属性列 y_i 的相对凝聚值为 $y'_i = \dfrac{\sum_{t=1}^{l} d_t s_t}{m}$。则用相对凝聚值表示的 D_1 为 $D_1 = \{y'_1, y'_2, \cdots, y'_n\}$。

第 3 步:用 K-均值算法对数据进行聚类。

(1) 初始时设前 k 个数作为簇的均值,并建立 k 个集合。依次分别把 D_1 中的前 k 个数据依次放入到 $S_1, S_2, \cdots, S_k (k \leqslant n, k \neq j)$ 中,即 $S_1 = \{y'_1\}, S_2 = \{y'_2\}, \cdots, S_k = \{y'_k\}$;

(2) 依次从 D_1 中取数据对象 $y'_{k+1}, \cdots, y'_n$,利用欧几里德公式计算每个数据对象与每个簇均值的距离 $d_{qh} = |y'_q - y'_h| (1 \leqslant q \leqslant k, k+1 \leqslant h \leqslant n)$,$d_{\min} = \min\{d_{qh}\}$;

(3) 则 $S_q = \{y'_q, y'_h\}$ 且 $y'_q = \dfrac{(y'_q + y'_k)}{2}$,再转第(2)步骤;

(4) 若这 k 个集合里面的元素不再发生变化时,聚类结束。

第 4 步:形成贝叶斯网络模型。

对聚类产生的各个元组中用鉴别信息来计算各个属性之间的相互关系。最后再增加类到各个元组之间的有向边。

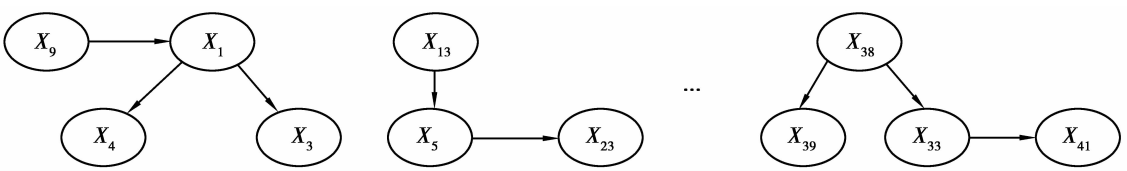

图 1 聚类后局部贝叶斯网络结构示意图

为了便于理解,设类属性为 C,各个非类属性为 $X = \{X_1, X_2 \cdots X_n\}$。设各个元组产生的聚类结果 $S_1 = \{X_1, X_3, X_4, X_9\}, S_2 = \{X_5, X_{13}, X_{23}\}, \cdots, S_k = \{X_{33}, X_{38}, X_{39}, X_{41}\}$。则经过第一步处理之后如图 1 所示。则最后形成的贝叶斯网络如 2 所示。

第 5 步:对缺损的值分类分析后进行修补。

设缺损的数据 x_{ij}所属列中含有 l 个分类,即:$C_1, C_2, \cdots, C_l$。贝叶斯公式 $P(C_i \mid X) = \dfrac{P(X \mid C_i)P(C_i)}{P(X)}$

中,令 $P(X \mid C_i) = \prod_{w=1}^{l} P(S_w \mid C_i)$,其中结构当中的局部贝叶斯网络的概率公式为 $P(X_1, X_2, \cdots, X_n) = \prod_{i=1}^{N} P(X_i \mid P(X_i))$。取 $P(C_h \mid x_j) = \max\{P(C_h \mid x_j)\}$,其中$(h = 1,2,\cdots,l)$,则此时将 x_{ij} 分类到 C_h 中。修复 x_{ij},令 $x_{ij} = \overline{x_{hj}}$,且 $x_{hj} \in C_h$。为了便于理解,在此给出此算法的流程图如图 3。

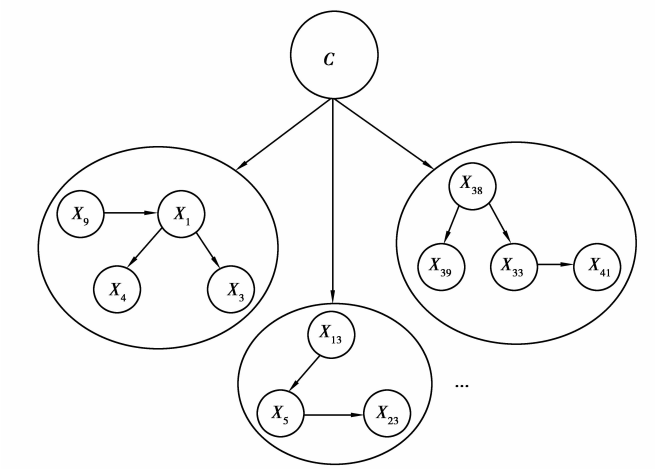

图 2　形成的贝叶斯网络结构示意图

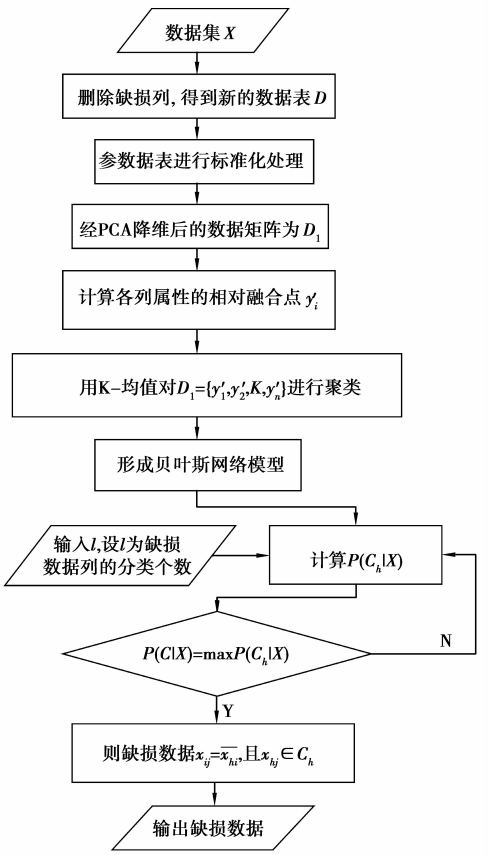

图 3　改进的朴素贝叶斯网络算法流程示意图

3　算法性能分析

分类算法模型是以在朴素贝叶斯网络的基础上提出来的,与之不同之处在于算法摒弃了非类属性变量相对于类属性变量是相对独立的前提条件。在改进的算法中,由于处理的数据非常的庞大,故在算法开始就借用了主成分分析法在对数据信息量保存的同时来进行降维操作,这样对算法着重于分类模型的研究有很大的帮助。在改进的算法中,给出了相对融合点这个概念,并给出了获取相对融合点的算法。最后算法利用 K-均值来研究各列属性之间存在的隐含的依赖关系,有利于将列属性值进行融合,这样就简化了下一步的 K-均值聚类算法在数据中的处于算法对缺损值的分类。算法相当于是扩大了朴素贝叶斯网络应用的领域。当对数据量很大,且各维属性之间又存在着隐含关系时,朴素贝叶斯网络很难达到理想的效果,而改进的算法将是处理这些问题行之有效的办法。

4　实验与分析

实验采用“某工厂 3 月份高炉数据”样本数据表作为数据集,共有 200 条数据记录,11 个属性:硅 Si,锰

Mn,磷 P,硫 S,钛 Ti,铬 Cr,镍 Ni,铁 Fe,锌 Zn,铜 Cu,镁 Mg;其中有几个缺失的数据,缺失数据会对分析结果造成一定的影响,减小结果的准确度,必须使用合适的方法对数据进行修补。该数据表的部分数据情况如下表1所示。

表1 某工厂3月份高炉数据数据表

硅 Si	锰 Mn	磷 P	硫 S	钛 Ti	铬 Cr	镍 Ni	铁 Fe
0.992	0.801	0.105	0.017	0.133	0.555	0.154	94.14
0.849	0.656	0.092	0.027	0.096	0.546	0.176	337.85
0.528	0.578	0.092	0.031	0.067	0.484	0.188	265.79
0.575		0.091	0.028	0.085	0.498	0.178	229.12
0.986	0.778	0.101	0.019	0.161	0.447	0.157	49.90
0.923	0.705	0.100	0.019	0.132	0.312	0.132	107.71

下面以缺失的第5条数据的“锰 Mn”属性值为例,通过实际的实验来说明如何修补该缺失的值。第一步首先删除该缺失值所在的属性列“锰 Mn”,对数据表进行标准化处理,那么该数据表变为一个200行10列的数据表,用矩阵表示如下:

$$D=\begin{bmatrix} x_{1,1} & x_{1,3} & x_{1,4} & x_{1,5} & \cdots & x_{1,11} \\ x_{2,1} & x_{2,3} & x_{2,4} & x_{2,5} & \cdots & x_{2,11} \\ \vdots & & \vdots & \vdots & & \vdots \\ x_{200,1} & x_{200,3} & x_{200,4} & x_{200,5} & \cdots & x_{200,11} \end{bmatrix}$$

然后对该数据表进行标准化处理,再通过 PCA 算法作用后,数据表由10维降到了8维,去掉了“钛 Ti”,“铜 Cu”。降维后的数据表用矩阵 D_1 表示如下:

$$D_1=\begin{bmatrix} x_{1,1} & x_{1,3} & x_{1,4} & x_{1,6} & x_{1,7} & x_{1,8} & x_{1,9} & x_{1,11} \\ x_{2,1} & x_{2,3} & x_{2,4} & x_{2,6} & x_{2,7} & x_{2,8} & x_{2,9} & x_{2,11} \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & \vdots \\ x_{200,1} & x_{200,3} & x_{200,4} & x_{200,6} & x_{200,7} & x_{200,8} & x_{200,9} & x_{200,11} \end{bmatrix}$$

求降维后各列属性值的相对融合点,并利用相对融合点结合 K-均值算法对数据进行聚类,聚类结果得到如下属性分组:$S_1=\{$铁 Fe$\}$,$S_2=\{$硅 Si,锌 Zn,镁 Mg$\}$,$S_3=\{$磷 P,硫 S$\}$,$S_4=\{$铬 Cr,镍 Ni$\}$。

对删除的属性列“锰 Mn”用数学家史特吉斯(Sturges)提出的公式:$k=1+3.32\log n$ 来对该列数据进行分组,其中 n 为数据集中数据的个数;则属性值 x_i 如果在$[\min+l*((\max-\min)/k),\min+(l+1)*((\max-\min)/k)]$区间内,则变换后的值为 l。其中 $l=0,1,\cdots,k$,min 是属性列中的最小值,而 max 是属性列中的最大值,这里每一个分组被认为一个类,有18个分类,即:$C_1,C_2,\cdots,C_{18}$。

最后对缺失的 $x_{5,2}$ 进行修补,具体如下:

计算 $P(C_i\mid X)=\dfrac{P(X\mid C_i)P(C_i)}{P(X)}$,其中 $P(X\mid C_i)=\prod_{w=1}^{18}P(S_w\mid C_i)$,取 $P(C_h\mid x_{5,2})=\max\{P(C_h\mid x_{5,2})\}(i=1,2,\cdots,18)$,则此时将 $x_{5,2}$ 分类到 C_h 中。然后可以取类 C_h 所在的区间的一个均值0.599作为 $x_{5,2}$ 的值进行修补。

5 结束语

针对朴素贝叶斯网络分类模型在处理高维大数据量时的效率偏低和准确率有待提高的问题,在朴素贝

叶斯网络的基础上结合主元分析法与 K-均值聚类算法构造出了一个改进的朴素贝叶斯网络的分类模型。该模型摒弃了非类属性变量相对于类属性变量是相对独立的前提条件,算法在一开始就用主元分析法在对数据集的信息量尽量保存的同时进行了降维操作,使得算法可以着重于进行分类问题。算法还提出了一个“相对融合点”的概念,并给出了相对融合点的获取方法,有效地提高了算法的性能。最后将改进的算法应用到质量管理中进行了实验,用算法产生的分类结果对数据集中产生的一些缺失数据进行修补,取得了较为理想的结果。

参考文献:

[1] PELIKAN M,GOLDBERG D,SASTRY K. Bayesian optimization algorithm,decision graphs,and Ocam's razor[R]. Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001),PP. 519-526. Also IlliGAL Report No. 2000020 (2001)

[2] FRIEDMAN N,GEIGER D,GOLDSZMIDT M. Bayesian Network Classifiers[J]. Machine Learning,1997,29:103-163

[3] PELIKAN M, SASTRY K, GOLDBERG D. Scalability of the Bayesian optimization algorithm [J]. International Journal of Approximate Reasoning,2002,31(3):221-258

[4] KAI M, ZHENG Z. A Study of AdaBoost with Naïve Bayesian Classifier: Weakness and Improvement [J]. Computational Intelligence,2003(19):186-200

[5] DING Z,PENG Y,PAN R. BayesOWL: Uncertainty Modeling in Semantic Web Ontologies[J]. In Soft Computing in Ontologies and Semantic Web,Springer-Verlag,December 2005

[6] HAN J,KAMBER M. Data Mining:Concepts and Techniques[M]. Academic Press,2001

[7] 王洪春,彭宏. 一种基于主成分分析的异常点挖掘方法[J]. 计算机科学:2007,10(34):192-194

A Naïve Bayesian Network Classification Model Based on K-means Clustering

LIU Ya-hui, WANG Yue, TAN Shu-qiu

(College of Computer Science and Engineering,Chongqing University of Technology,Chongqing 400054,China)

Abstract: According to the low efficiency and low accuracy of the naive Bayesian network classification model in dealing with large number of high-dimensional data,by combining Principal Component Analysis and K-means clustering algorithm, this paper gives an improved Naïve Bayesian network classification model. The model abandoned the premise for the relative independence between non-class attribute variables and class attribute variables. Firstly,we use principal component analysis to reduce the dimensionality of the data set,so the algorithm can focus on the classification problem. The algorithm has also proposed a concept called “relative fusion point” to effectively improve the performance of the algorithm. Finally,the performance of the algorithm is analyzed,and the improved algorithm is applied to the actual data set for experiment to repair the missing data of the data set,the results show that the algorithm is effective.

Key words: Bayesian network classification;Naïve Bayesian network;K-means clustering;data mining

责任编辑:代小红