基于 Isomap 降维的噪声处理算法①

屈治礼

(江苏科技大学 计算机科学与工程学院, 镇江 212003)

摘 要: 由于非线性降维方法对高维数据中存在的噪声比较敏感, 导致最终的分类效果比较差. 为了弥补其不足, 在首先使用极大似然估计方法估测出样本数据本征维度的前提下, 提出一种结合等距特征映射与主成分分析的方法. 一方面能够使原始数据保持其在高维空间的几何结构, 另一方面可以消除噪声对降维结果的影响, 最终使得低维数据尽可能的保持原始样本数据集的内在特征. 通过实验论证表明, 该组合方法的效果比单独直接使用等距特征映射和主成分分析算法的效果都要好.

关键词: 等距特征映射; 极大似然估计; 高维数据; 噪声

Noise Processing Algorithm Based on Isomap Reducing Dimensionality

QU Zhi-Li

(College of Computer Science and Engineering, Jiangsu University of science and Technology, Zhenjiang 212003, China)

Abstract: Nonlinear dimensionality reduction method is more sensitive to the noise in the high-dimensional data, resulting in relatively poor final results of classification. In order to make up for its shortcomings, this paper proposes a method in the premise of using the maximum likelihood estimation method to estimate the intrinsic dimension of the sample data, which combines the isomapetric mapping with the principal component analysis. On the one hand, the method enables the original data to maintain its geometry in the high dimensional space, on the other hand, the method can eliminate the influence of noise on the dimensionality reduction results, eventually making the low-dimensional data as much as possible to maintain inherent characteristics of the original sample data sets. Experimental demonstrations show that the results of combination method is better than separate isometric mapping and separate principal component analysis.

Key words: isomapetric mapping; maximum likelihood estimation; high-dimensional data; noise

随着计算机技术在各行各业中的迅猛发展, 高维数据在我们的生活中可以频繁遇到, web 的文本和多媒体数据随处可见. 这种高维形式数据的存在, 已经超越了人们的认知能力. 如何将三维空间以上的多维或高维数据转化为人类能视觉直观理解的可视化结果, 是高维数据可视化(High Dimensional Data Visualization, 简称 HDDV)所研究的课题. 目前学者们提出了相应的非线性降维方法, 主要有等距特征映射(Isomapetric Mapping)[1,2]、局部线性嵌入(Locally Linear Embedding)[3]和拉普拉斯特征映射(Laplacian Eigenmaps)[4]等, 并在某些问题的处理上取得了较好的效果.

目前, 针对非线性降维算法的降维维数一般通过残差[1]来描述的不足, 本文利用极大似然估计(Maximum Likelihood Estimation, 简称 MLE)方法估测高维数据的本征维数(Intrinsic Dimension, 简称 ID); 针对非线性降维方法存在的对噪声比较敏感的问题, 张振跃[5]等提出局部线性平滑的思想, 采用加权来构建局部小块, 并用迭代方法优化权值, 但是这种迭代方法存在容易陷入局部极小值, 鲁棒性不够高等问题. 本文首先采用 MLE 方法估测出原始高维数据的本征维数, 其次使用 Isomap 算法将高维数据降维至本征维度的低维空间, 这样可以保持以前的非线性结构; 最

① 收稿时间:2013-04-12;收到修改稿时间:2013-05-20

后采用主成分分析 PCA[6-8]方法降维，以达到剔除一些无用噪声信息的目的. 通过实验证明了该组合算法 Mp_Isomap 的可行性与时效性.

1 等距特征映射

等距特征映射(Isomap)是由 Tenenbaum 等在 2000 年 Science 提出的一种非线性降维方法，它以经典多维尺度(Classical Multi-Dimensional Scaling，简称 CMDS)变换为基础，核心是保持两点间的测地距离(geodesic)，即把原始空间中距离的计算从欧氏距离变成了流形上的测地距离. Isomap 通过将数据点连接起来构成一个邻接图来离散地近似原来的流形，所谓流形(manifold)就是一般的几何对象的总称，流形就包括各种维数的曲线曲面等. 而测地距离也相应地通过图上的最短路径来近似了，使得距离很近的点间的测地距离用欧氏距离代替，距离较远的点间的测地距离用最短路径来逼近，这里测地距离是指在包含待测两点内的地球面上测得的两点之间的最短线. 其算法简单描述如下:

Step1: 生成邻域图 G. 计算任意两个样本 x_i 与 x_j 的欧式距离 $d_x(x_i, x_j)$，如果 x_j 在 x_i 的半径 ε 之内或者是 x_i 的 k 个最近邻点之一，则连接 x_j 和 x_j，连接线的长度设置为 $d_x(x_i, x_j)$，否则将 x_i 与 x_j 的长度置为 ∞； 其中 $1 \le i \le n$，$1 \le j \le n$，如此重复下去，构造无向图 G;

Step2: 计算任意两样本之间的最短距离. 利用迪杰斯特拉(Dijkstra)或弗洛伊德(Floyd)算法计算邻域图 G 中任意两样本的最短距离 $d_G(x_i, x_j)$，这样得到最短距离矩阵 $M_G = \{d_G(x_i, x_j)\}$;

Step3: 构建 d 维欧几里得空间嵌入. 将 CMDS 应用于下式距离矩阵，它由最短距离的平方组成:

$$D_G = M_G^2 = \{d_G^2(x_i, x_j)\}$$

下图 1 是瑞士卷经过 Isomap 过程得到的二维嵌入结果图 2.

对一个光滑流行采样后获得样本数据，使用一些非线性降维方法可以找到其本征维数，但是，由于在采样过程中不可能保证在理想状态下，这样就会带来噪声的干扰，使得降维至低维空间时出现对原始数据的变形. 下图 4 表明在图 3 曲线中加入信噪比 SNR(Signal to Noise Ratio)=10dB 的 Gauss 白噪声后，结果呈现了严重的扭曲，如图 5.

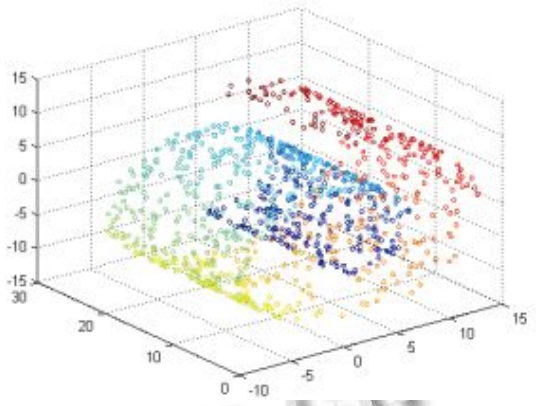

图 1 Swiss roll 原始数据集

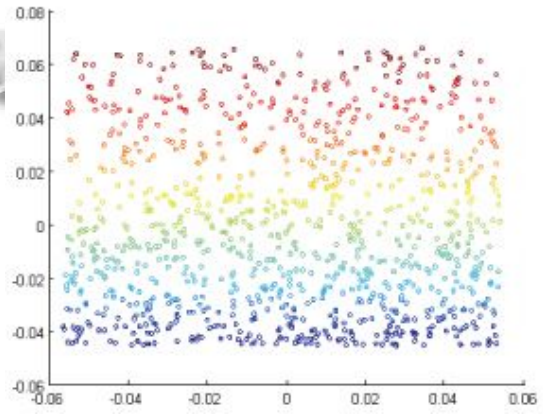

图 2 Isomap 二维嵌入结果

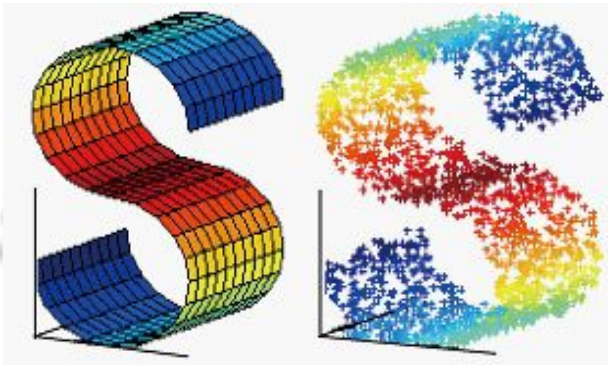

图 3 原始曲线 图 4 加入高斯白噪声的曲线采样点

图 5 使用 Isomap 方法的低维射结果

2　极大似然估计本征维数

极大似然估计法是建立在极大似然原理的基础上的一个统计方法. 对观测数据集进行建模所需的最少独立变量的个数, 通常称之为最优嵌入维数, 也称为本征维数[3,9], 虽然上述残差图可以近似估计高维数据的本征维数, 但是这种观察得到的值依赖于人为的观察, 且邻域大小难以选取、计算耗时, 不利于我们对本征维数的估测. 以下我们采用极大似然估计方法来对高维数据的本征维数加以测度.

假设高维空间 R^{p} 中有 n 样本 $X_1, \mathrm{X}_2, \cdots, \mathrm{X}_n$, 可在低维空间 $\mathrm{R}^{\mathrm{m}}\left(\mathrm{m} << p\right)$ 中用 $Y_1, Y_2, \cdots, Y_n$ 嵌入表示, 即: $X_i = g(Y_i)$, $i = 1,2,\cdots,n$, g 是一个连续并且足够光滑的嵌入映射. 其中 m 称为本征维数. 假设对于给定的一个点 x, R 的足够小范围内, 以其为半径的球体 $S_x(R)$ 内 $f(x) \approx$ 定值. 先来看下面这个过程:

$$\begin{aligned} &\{N(t,x), 0 \le t \le R\} \\ &N(t,x) = \sum_{i=1}^{n} I\{X_i \in S_x(t)\} \end{aligned} \tag{1}$$

观察可知 $N(t,x)$ 是 $X_1, \mathrm{X}_2, \cdots, \mathrm{X}_n$ 落入 $S_x(t)$ 的点数, $N(t,x) = \sum_{i=1}^{n} I\{X_i \in S_x(t)\}$ 可近似为泊松过程. 注意到, 对于 $X_1, \mathrm{X}_2, \cdots, \mathrm{X}_n$, 记 $T_k(x)$ 为 $X_1, \mathrm{X}_2, \cdots, \mathrm{X}_n$ 中 x 的第 k 个近邻(从近到远)到 x 的距离, 则有:

$$\frac{k}{n} \approx f(x)V(d)\left[T_k(x)\right]^d \tag{2}$$

其中 $V(d) = \pi^{d/2}\left[\Gamma(d/2)+1\right]^{-1}$ 指的是 d 维单位球体的体积.进一步假定 t 固定, 可得 $\lambda(t) = f(x)V(d)dt^{d-1}$, 考虑到 $N(t,x)$ 对 x 的独立, 易知 $\lambda(t)$ 是 $N(t)$ 相对于 t 的变化率. 记作 $\theta = Inf(x)$, 在这里, 根据文献[1]对泊松过程的建立的对数似然函数:

$$InL(d,\theta) = In\left(\int_0^R In\lambda(t)dN(t) - \int_0^R \lambda(t)dt\right) \tag{3}$$

其满足下式似然方程:

$$\frac{\partial InL}{\partial \theta} = \frac{1}{\int_0^R In\lambda(t)dN(t) - \int_0^R \lambda(t)dt}\left(\int_0^R dN(t) - \int_0^R \lambda(t)dt\right) = 0$$

化简即:

$$\frac{\partial InL}{\partial \theta} = \frac{1}{\int_0^R In\lambda(t)dN(t) - \int_0^R \lambda(t)dt}\left(N(R) - e^{\theta}V(d)R^d\right) = 0$$

由 $\int_0^R In\lambda(t)dN(t) - \int_0^R \lambda(t)dt \ne 0$, 则:

$$N(R) - e^{\theta}V(d)R^d = 0 \tag{4}$$

$$\frac{\partial InL}{\partial d} = \frac{\left(\frac{1}{d} + \frac{V'(d)}{V(d)}\right)W(R) + \int_0^R IntdW(t)}{\int_0^R In\lambda(t)dW(t) - \int_0^R \lambda(t)dt} - \frac{e^{\theta}V(d)R^d\left(InR + \frac{V'(d)}{V(d)}\right)}{\int_0^R In\lambda(t)dW(t) - \int_0^R \lambda(t)dt} = 0$$

由 $\int_0^R In\lambda(t)dN(t) - \int_0^R \lambda(t)dt \ne 0$, 则:

$$\begin{aligned} &\Rightarrow \left(\frac{1}{d} + \frac{V'(d)}{V(d)}\right)W(R) + \int_0^R IntdW(t) - \\ &\int_0^R IntdW(t) - e^{\theta}V(d)R^d\left(InR + \frac{V'(d)}{V(d)}\right) = 0 \end{aligned} \tag{5}$$

综合式(4)和式(5)可得:

$$\hat{d}_R(x) = \left[\frac{1}{N(R,x)}\sum_{j=1}^{N(R,x)} In\frac{R}{T_j(x)}\right]^{-1} \tag{6}$$

由于取球形邻域并不太方便我们的计算, 取 k 邻近更易操作, 此时(6)变为下式(7):

$$\hat{d}_k(x) = \left[\frac{1}{k-1}\sum_{j=1}^{k-1} In\frac{T_k(x)}{T_j(x)}\right]^{-1} \tag{7}$$

对于 k 固定, 将上式中的 x 遍历样本 $X_1, \mathrm{X}_2, \cdots, \mathrm{X}_n$, 得到的本征维数并不是很理想的; 如果让 k 在一定范围内取值, 将每一次取定 k 后得到的估计值再求一次平均值, 这样最终得到的本征维数的估计值的可信度会更高.

$$\hat{d}_k = \arg\left(\hat{d}_k(X_i)\right) \tag{8}$$

基于上述过程, 假定介于 0 和 100 之间. 现选用数据集分别是: Swiss roll[3], Twin peaks 为人造数据集, Faces 数据集[3]和 Hands rotation 序列[10].

采用极大似然估计 MLE 的式(8)估计出本征维数, 并与残差法作相应的比较, 横轴表示近邻点个数, 纵轴表示本征维数, 如下图 6 到图 9 所示.

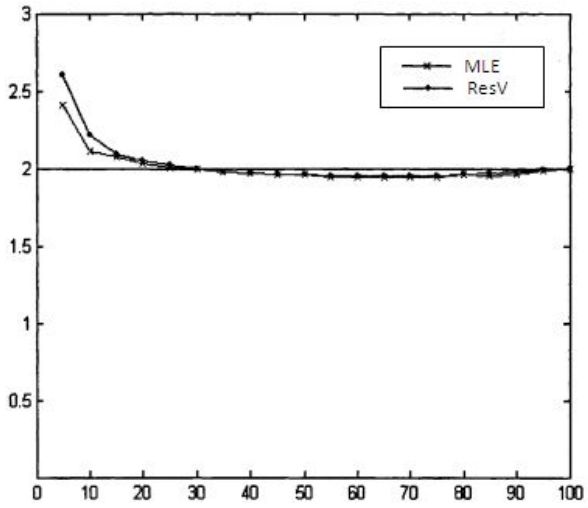

图 6 Swiss roll 上的残差和 MLE 比较

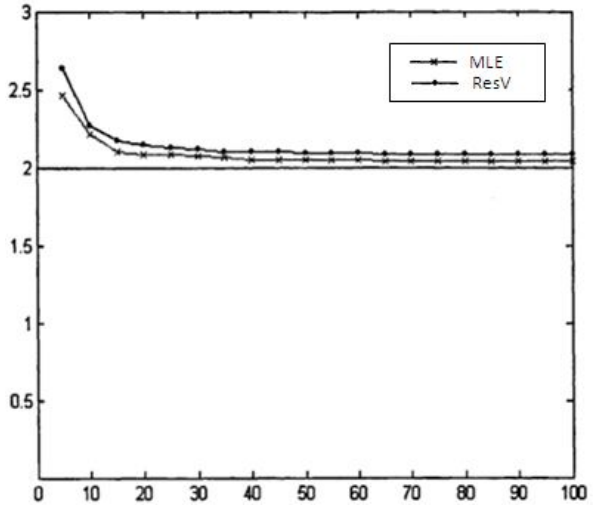

图 7 Twin peaks 上的残差和 MLE 比较

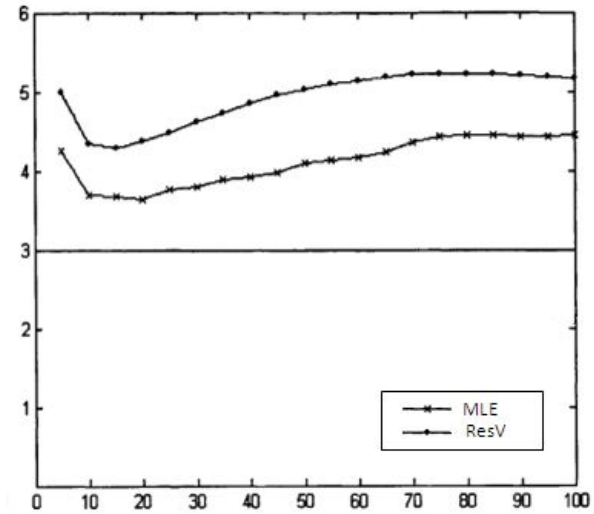

图 8 Faces 上的残差和 MLE 比较

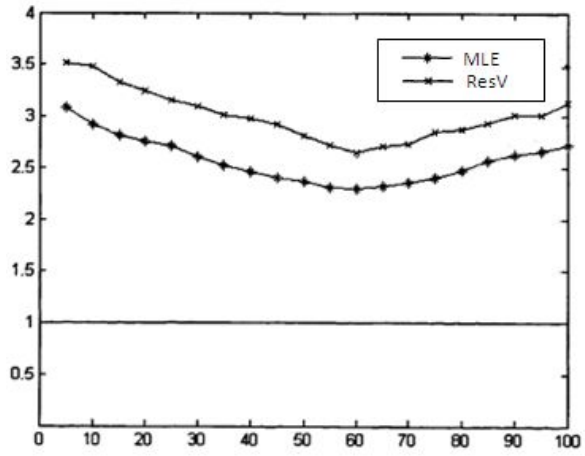

图 9 Hands rotation 上的残差和 MLE 比较

将结果求得平均值后, 汇总如下表 1 所示:

表 1 MLE 与残差估计本征维数的比较

数据集	样本	数据维数	本征维数	MLE	残差法
Swiss roll	2000	3	2	2.05	2.1
Twin peaks	1000	3	2	2.1	2.2
Faces	698	64×64	3	3.5	4.6
Hands rotation	481	480×512	1	2.5	3.1

从表中数据我们可以看到, MLE 对高维数据的本征维数有很好的估测作用; 对于已知的数据集本身而言, MLE 比残差法对本征维数的估计更接近于实际数据本身.

3 Mp_Isomap流程和实验应用

对噪声处理的算法 Mp_Isomap 实现流程归纳如下:

Step1: 使用 MLE 估计原始样本的本征维数 d, 执行 Isomap 算法, 构建 d 维空间嵌入;

Step2: 使用 PCA 算法降维, 过滤无用噪声信息;

Step3: 调整 Isomap 的近邻数 K, 保证其准确率.

实验采用 UCI 数据库中的手写数字来做实验. 这个数据库中包含了训练样本集和测试样本集. 训练样本集中共有 3823 个样本, 测试样本集中共有 1797 个样本, 每个样本是一个 65 维的向量, 其中最后一维表示该样本的类别. 这里, 取训练样本集中的 1500 个样本用作训练, 取测试样本集中的 1000 个样本作为测试. 采用最近邻方法进行分类. 这里, 设定近邻数 k=5, 6, …, 10.

表 2　Isomap 的平均错误率

维数d/ 平均错误率	3	4	5	6	7	8	9
Isomap	0.096	0.074	0.058	0.056	0.064	0.062	0.056
PCA	0.204	0.232	0.056	0.118	0.104	0.086	0.074
Isomap+PCA	0.094	0.072	0.052	0.052	0.048	0.048	0.052

表 3　PCA 的平均错误率

维数 d	3	4	5	6	7	8	9
平均 错误率	0.204	0.232	0.176	0.118	0.104	0.086	0.074

首先使用 MLE 得到本征维数 d=16, 如图 10 所示:

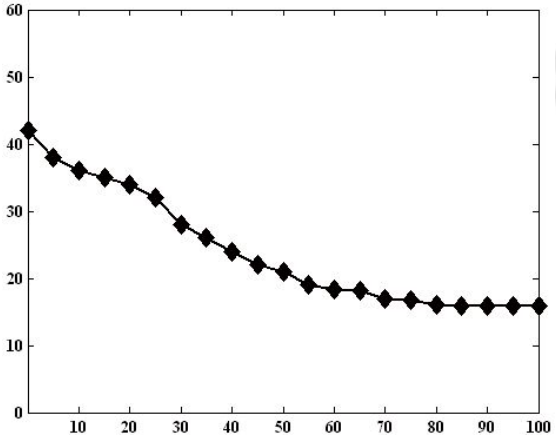

图 10　手写数字的 MLE 估测本征维数

这里, 我们用 Isomap 方法首先将原始数据降到 16 维, 此时的分类错误率如下:

表 4　Isomap 的平均错误率

近邻 数 k	5	6	7	8	9	10
平均 错误率	0.042	0.044	0.048	0.054	0.052	0.060

其次, 用 PCA 方法将维数分别降低到 d=3、4、5、6、7、8、9 维, 则此时的分类平均错误率如下表 5 所示.

表 5　PCA 的平均错误率

维数d/ 近邻数k	3	4	5	6	7	8	9
5	0.094	0.072	0.052	0.056	0.048	0.052	0.052
6	0.126	0.086	0.054	0.052	0.052	0.048	0.050
7	0.132	0.102	0.082	0.082	0.058	0.058	0.054
8	0.170	0.094	0.086	0.072	0.074	0.062	0.540
9	0.142	0.096	0.070	0.074	0.076	0.066	0.064
10	0.178	0.134	0.114	0.092	0.086	0.082	0.082

最后总结该过程三种方法的平均错误率如表 6 所示.

表 6　三种算法分类的比较

维数d/ 平均错误率	3	4	5	6	7	8	9
Isomap	0.096	0.074	0.058	0.056	0.064	0.062	0.056
PCA	0.204	0.232	0.056	0.118	0.104	0.086	0.074
Isomap+PCA	0.094	0.072	0.052	0.052	0.048	0.048	0.052

从上述实验结果容易发现, 使用 Isomap +PCA 的方法, 所得到的平均错误率要比直接使用 PCA 要低很多, 相对直接使用 Isomap 而言, 分类效果也要好一点. 对这个数据样本集来说, 我们可以观察, 当近邻数取 k=7 或 k=8 时, Isomap +PCA 方法得到的效果是比较好的. 这也说明了这种组合的方法是具有一定的优越性的.

4　结语

从上述实验的结果来看, 这种 Isomap 和 PCA 组合的方法 Mp_Isomap 所获得的分类效果还是比较理想的, 它的平均错误率既低于直接的 PCA 方法, 又低于原始的 Isomap 算法. 针对极大似然估计方法, 由于整个数据集的估计维数是每个点估计维数的平均值, 在估测过程中, 有偏离本征维数误差较大的情况出现; 这里给出的近邻点是人为给定的范围, 对于任意未知的数据集, 如何确定一个合适的近邻点也是比较困难的; 以上两点将是下一步进行研究的方向和改进的地方.

参考文献

1 Tenenbaum JB, Silva V, Landford JC. A global geometric framework for nonlinear dimensionality reduction. Science, 2000, 290(22): 2319–2323.

2 余肖生,周宁.高维数据降维方法研究.情报科学,2007,25(8): 1250–1251.

3 Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000, 290 (5500): 2323–2326.

4 Belkin M, Niyogi P. Laplacian eigenmaps for dimension-nality reduction and data representation. Neural Computation, 2003, 15(6): 1373–1396.

5 Park JH, Zhang ZY, Zha HY. Local linear smoothing for nonlinear manifold learning. IEEE Computer Society Conference on CVPR. Washington DC. 2004, 2:452–459.

6 Telling H. Analysis of a complex statistical variable into

(下转第 94 页)

确定中断原因, 解决完错误后才能重启接收程序.

发送程序流程图如图 6 所示.

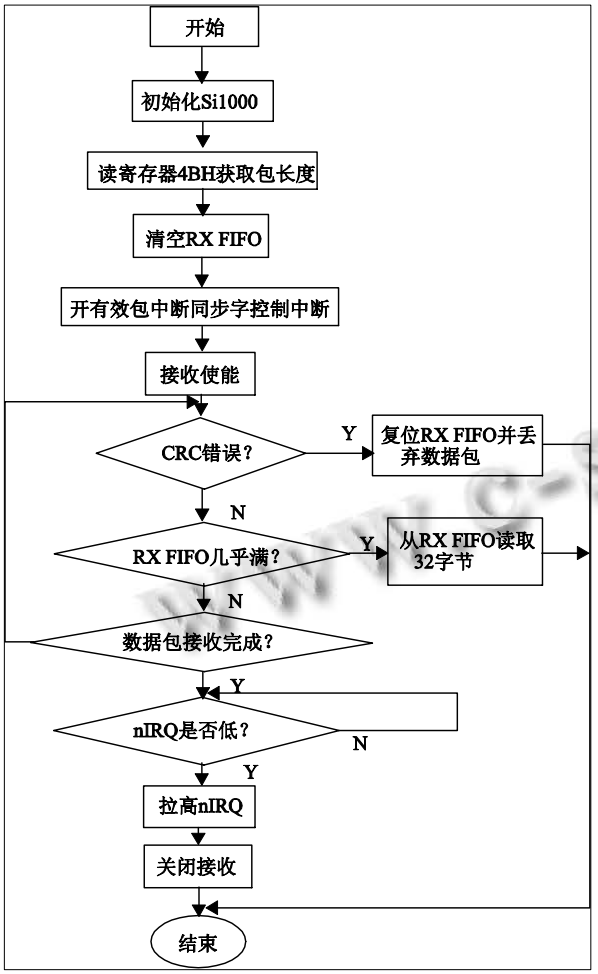

图 6 接收程序流程图

4 结论与展望

Si1000 是一款微功耗、体积小并有强大无线通信功能的微控制器, 用它作为核心元件采集数据, 进行从站与主站的无线通信可靠性高. 从站与主站间的传输距离过长时可在中间加设集中器, 从而保证正常的接收与发送数据. Si1000 为 MCU 可以采集现场各种水、暖、电表的数据, 并经过无线传输到监控平台, 大大简化了由布线带来的困难. 此外, 将 Si1000 作为无线传感网络的普通节点形成无线自组织网络, 应用到安防、无线抄表、智能家居、环境医疗检测等领域, 为无线通信提供了更加可靠的硬件平台.

参考文献

1 Lin CE, Li CC, Hou AS. A real-time remote control architecture using mobile communication. IEEE Transactions on Instrumentation and Measurement, 2003, 52(4):997–1003.

2 Silicon Laboratories.Ezradiopro Programmingfuide. 2009. http://www.silabs.com/support/pages/support.aspx Product Family = EZRadioPRO.16–45.

3 李善荣,闫述.Si1000 低功耗性能与在无线传感器节点上的应用开发.无线通信技术,2011(3):32–37.

4 黄智伟,李富英.基于射频收发芯片 nRF401 的计算机接口电路设计.微电子学与计算机,2002(5):40–41.

5 单海东,卢东贵.基于 Si1000 无线微控制器的无线射频测. 自动化系统工程,2010(8):120–122.

6 李伯成.单片机及嵌入式系统.北京:清华大学出版社, 2008:259–363.

7 Silicon Labs Ezmac and Ezhop User's Guid Edatasheet.2010.

8 Silicon Laboratories.Si1000 /1 /2 /3 MCU with Integrated 240 − 960 MHz EZRadioPRO Transceiver. 2010. 93–99, 122–130, 200–245, 248–265.

(上接第 114 页)

principal components. Journal of Educational Psychology, 1933, 24: 417–441.

7 Turk Mand Pentland A. Eigcnfaces for recognition. J. Cognitive Neuroscicnce, 1991, 3(1): 71–86.

8 雷君虎,杨家红,钟坚成等.基于 PCA 和平行坐标的高维数据可视.计算机工程.2011,37(1):48–50.

9 Verleyse M. Learning high-dimension data. Limitations and Future Trends in Neural Computation. Amsterdam. The Netherlands, IOS Press. 2003. 141–162.

10 Kegl B. Intrinsic dimension estimation using packing numbers.http://books.nips.cc/papers /files/nips15/AA25.pdf.