Probability of large-scale data set EM clustering algorithms based on partial information constraints

Xiaoyan Liu[1,a]

[1]Changchun University of Science and Technology, Changchun City of Jilin Province, 130600, China

[a]xiaoyan2008066@163.com

Keywords: Some constraint information; Clustering; The data set; The clustering quality; The probability of clustering algorithm

Abstract. The current situation, the need for clustering of data is very large, and the use of traditional algorithm for clustering process often tedious and time consuming is very long, the effect is not obvious. Based on this, this paper proposes a data sets EM probability based on some constraint information clustering algorithm, the detailed implementation process of the whole algorithm is described. Through experiment contrast scalable EM, positive_PC_SEM and full_PC_SEM clustering quality and efficiency of execution of the algorithm, the results show that the positive_PC_SEM algorithm and scalable EM algorithm compared to the clustering quality and efficiency is higher, although full_PC_SEM clustering quality is very high, but requires a lot of time.

Introduction

In order to cope with the need of efficient clustering of large scale data sets, a lot of research work has been done in the traditional deterministic clustering. But these algorithms because of their attribution to determine the standard is very strict or there are some defects, as a data and can only belong to a category [1]. However, in reality, there will be an object of the probability which belonging to several categories of the situation [2]. For example: a sports enthusiast, he may be a regular basketball enthusiasts to participate in basketball, but also may be a regular football fans to play football. Therefore, it is necessary to use probabilistic clustering method to get the probability of a particular object belonging to a class.

Clustering can be used as a density estimation process, the method of data generation can be used to express probability density function, and the mixed model can be used to express probability density function [3]. If the data can only be generated by a density function which belongs to the spherical Gauss distribution, and is unique, the clustering process is the traditional clustering process. If the limited relaxation, data that is not by one but can be composed of multiple Gaussian distribution density function according to a certain probability is generated. At the same time, density function is not required must be spherical Gaussian distribution, thus the clustering process and become a mixture of probabilistic clustering model.

EM algorithm [4] is currently the application of clustering technology is very extensive and very effective, estimating mixture model parameters of the algorithm, and the parameter gradually improve the model, finally terminating in a maximum point. Based on the density estimation theory, the mixed Gauss model can be used to express any data distribution, so it is usually used to take the mixed Gauss model in the implementation of clustering using EM algorithm.

At present, clustering data sets very much, using classical clustering algorithm, the process is often tedious and time-consuming and can‘t meet the actual demand. And according to the concept of mixed model, the importance of that data is not the same, can be divided into the following three categories: Statistics and the results are stored data can be deleted; preserved statistical information after data compression; for model calculation must be stored data. According to this idea, Microsoft proposed Scalable EM algorithm [5], this algorithm only need to scan disk data sets can be calculated raw data generated by the model. Compared with the traditional algorithm, the efficiency of the algorithm is much higher.

PC_SEM Algorithm

Using EM algorithm to deal with the current increasingly large scale data set, its efficiency is very low. The new algorithm is implemented in a single scan can complete probabilistic clustering is present many scholars research direction. The distribution of the original data set and initialization parameters are important factors that affect the probability of clustering algorithm. At present, these two factors have a significant impact on the probability clustering algorithm, which makes the clustering result is not stable and the quality is not high.

For the above reasons, this paper proposes a PC_SEM algorithm based on semi supervised clustering for the first time, which is a probabilistic clustering algorithm for large scale data sets based on partial constraint information. Data sets can automatically get partial constraint information, and the algorithm uses this information to guide the whole clustering process, so as to improve the quality of clustering results and the efficiency of clustering [6].

Rationally using the concept of a semi-supervised clustering, to a large extent can accelerate the convergence rate of the clustering process and improve the quality of clustering results, make the efficiency of clustering significantly increased. PC_SEM algorithm is proposed in this paper, based on the part of the constraints of the large-scale data set probability of EM clustering algorithm.

Can use the following procedure to describe the basic idea of PC_SEM:

Initialization for clustering of data set $R^0=\varnothing$,To discard a triple $B^0=\varnothing$ sets the main compression ,Set $C^0=\varnothing$ for triple time compression ,Setup has made a number of data points of processing $m^0=0$,Select a particular data set D way of reading, can random back sampling can also be in order. For each read data sets$S^T\subset D$, all need to set PC_SEM clustering main bufferSize, setting ε limit for PC_SEM stop calculation. The following are specific steps:

Step 1: To read the data sets, its available bufferSize according to the PC_SEM, $R^{T+1}=R^T\cup S^T$, $R^{T+1}\cup C^T\cup B^T=\text{bufferSize}$, $m^{T+1}=m^T+|S^T|$.

Step 2: According to $R^{T+1}\cup C^T\cup B^T$ perform extended, replace with ϕ^{T+1} to ϕ^T .

Step 3: To distinguish the processing of data from a data set R^{T+1}, identify belonging to B^T point and the data points belonging to C^T sets, the original data set using a triple to replace, and remove the R^{T+1}related set of data points. In this way can compress the data point set, through the implementation of main memory space to clear, can provide convenient for subsequent read in data collection. Note, do not need to remove the constraint information.

Step 4: Termination conditions:when $T>1$ and $\overline{L}(\phi^T,m^T)-\overline{L}(\phi^{T+1},m^{T+1})\le\varepsilon$,Stop PC_SEM algorithm (If an early end PC_SEM algorithm, suggests that already find out all the data model. when $m^{T+1}=m^T$, said had processed the whole data set, PC_SEM algorithm has been terminated.

).Otherwise, $T\leftarrow T+1$, skip to step 1, $\overline{L}(\phi^T,m^T)=\frac{1}{|m^T|}\sum_{x\in m^T}\log(\sum_{h=1}^{k}w_h\cdot h_h(x\,|\,\mu_h,\Sigma_h))$.

In order to better complete the clustering process, expand PC_SEM algorithm in the implementation of the process, the need to read part equivalent to the relevant constraints information. Perform extended PC_SEM process description is as follows: According to $R^{T+1}\cup C^T\cup B^T$,initialization mixture model parameters ϕ^0 , ϕ^t replacement for ϕ^{t+1} in the iteration process. The following are specific steps:

Step 1: For R^{T+1} set for related parts of the constraint information pcR^- and pcR^+, pcR^- says some negative constraint information, pcR^+ said some constraint information.

Step 2: Initialization parameter μ_h , Σ_h , w_h , $h=1,\cdots,k$ .in which $\mu_h=Mean(x_i)$, $\Sigma_h-Cov(x_i)$, $x_i^h\in R^{T+1}$, $x_i^h\in pcR^+$, $w_h=\frac{1}{k}$.

Step 3: For every little classification in R^{T+1}, calculate the probability of it belongs to the clustering $h=1,\cdots,k$.

$$p(Y_j = h \mid X_j, \theta^{old}) = \frac{w_h^{old} \prod_{x_i \in X_j} p(x_i \mid y_i^j = h, \theta^{old})}{\Sigma_{m=1}^{k} w_m^{old} \prod_{x_i \in X_j} p(x_i \mid y_i^j = m, \theta^{old})}, \{X_j\}_{j=1}^{L}, L \le N 。$$

$$w_h(x) = p(Y_j = h \mid X_j, \theta^{old}) = \frac{w_h^{old} \prod_{x_i \in X_j} p(x_i \mid y_i^j = h, \theta^{old})}{\Sigma_{m=1}^{k} w_m^{old} \prod_{x_i \in X_j} p(x_i \mid y_i^j = m, \theta^{old})}, \quad \{X_j\}_{j=1}^{L}, L \le N, \quad x \in X_j 。$$

Step 4: Calculation of triples represented data points belonging to the probability of KKKK respectively.

$$\underset{(\theta,\Gamma,N)\in C^T \cup B^T}{w_h^t(\theta, \Gamma, N)} = \frac{w_h^t \cdot f_h(\frac{1}{N}\theta \mid \mu_h^t, \Sigma_h^t)}{\sum_{i=1}^{k} w_h^t \cdot f_h(\frac{1}{N}\theta \mid \mu_h^t, \Sigma_h^t)}$$

Step 5: Updated mixed model parameters ϕ^{t+1}

$$N(h) = \sum_{X_j \in R^{T+1}} p(Y_j = h \mid X_j, \theta^{old}) + \sum_{(\theta,\Gamma,N)\in C^T \cup B^T} n \cdot w_h^t(\theta, \Gamma, N), \quad h = 1,\mathrm{L}\ \mathrm{L}\ ,k, \quad N = \Sigma_{h=1}^{k} N(h) 。$$

Fully considering the parts of the constraint information and triples, assuming $w_{positive}^{h(t+1)} = \frac{N(h)}{N}$,To read the pcR^- calculation $w_{anti}^{h(t+1)}$ at the same time, this is part of negative constraint information for reasonable utilization, with $w_{anti}^{h(t+1)}$ has carried on the correction to $w_{positive}^{h(t+1)}$.

Step 6: Termination conditions: when $\left|\hat{L}(\phi^{t+1}, R^{T+1}, C^T, B^T) - \hat{L}(\phi^t, R^{T+1}, C^T, B^T)\right| \le \varepsilon$ Stop algorithm .Otherwise, skip to step 3.

The experimental results and discussion

In this paper, the real data sets using the Matlab language was designed and implemented respectively scalable EM and PC_SEM algorithms for large-scale data sets to test PC_SEM EM clustering validity. Real data set with KDD said it recorded the related situation of charitable giving. Article of 84513, any record data set and each record is 481 dimensions, select 56 dimensions to implement clustering.

Hypothesis point data all dimensions have the same weight, that each dimension of the importance of the same, so we need to standardize each dimension: $normal_i = \frac{data_i - mean_i}{\sqrt{(sumsq_i / num) - (sum_i / num)^2}}$, $i = 1, 2, \mathrm{L}\ , n$ .Because of the data of real category is not clear, so experiment with the method of using the manual annotation. In this paper, the kmeans algorithm is used for 50 times clustering, according to the actual situation of clustering generated some constraint information. In measuring the clustering quality, this article adopts the method of average logarithmic likelihood.

Assuming that main memory can accommodate a maximum of 6000 pieces of information, discard%=30%,set the threshold $\varepsilon = 1e-6$ .In the process of every cycle, use noc selection to represent the constraints of cycles, with a certain probability to choose some negative information. The original data set may effect on the clustering quality, in order to avoid the clustering results to a great extent quality affected, adopt the method of disrupted the original data.

Experiment was carried out on the original data in the process of three random, for each time USES the scalable EM, positive_PC_SEM and full_PC_SEM operations, 10 times each algorithm in clustering, a total of 30 times. Document clustering quality test results as shown in table 1.

Table .1 Three algorithms of clustering quality comparison results table

scalable EM	positive_PC_SEM	full_PC_SEM
-321.7	-107.6	-81.5

Analysis of table 1 can be seen that, in view of the real data sets, can also be good to restrain some information into the algorithm, which to a great extent, improve the quality of the clustering results.

To see the execution efficiency of each algorithm, as shown in figure 1 are averages of record. Can be seen from the graph, the algorithm of scalable EM and positive_PC_SEM algorithm, compared with the former clustering efficiency and quality are better than the latter. It shows that after the part of the constraint information integrated into can greatly accelerate the convergence speed of the algorithm. Full_PC_SEM algorithm execution efficiency is low, this is mainly because the algorithm utilizes the negative constraint information, consumes a lot of time in the implementation process, but its clustering is one of the best quality.

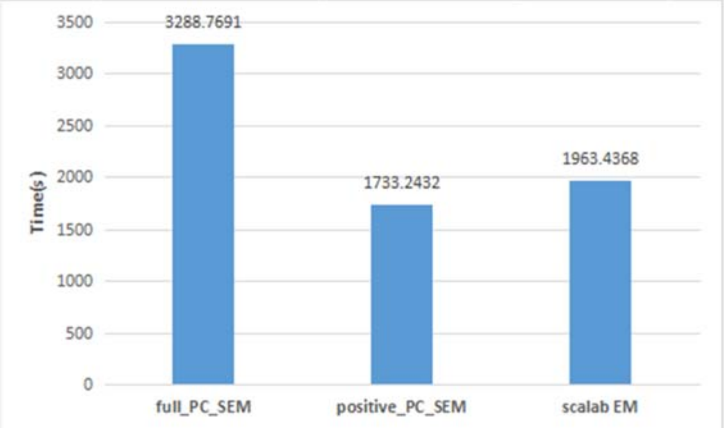

Figure .1 All kinds of average algorithm execution time

Conclusion

This paper puts forward a new algorithm of PC_SEM. This new algorithm in the process of clustering part constraint information can be adopted to carry out optimization, can be achieved by using the method of iteration in the limited implementation of data aggregation class work within main memory space. This article detailed part explains how the constraint information in large-scale data set probability of EM clustering process applications.

The experimental results show that compared with the algorithm of scalable EM, positive_PC_SEM algorithm clustering quality and efficiency is higher. Full_PC_SEM algorithm compared with positive_PC_SEM algorithm, it adopted some negative constraint information, so the algorithm of clustering results better quality. At the same time, time also will increase, the algorithm for clustering results have special requirements of application of the algorithm is more appropriate.

References

[1] ShenYan. Large-scale data set and efficient data mining algorithm, clustering algorithm [D]. The study of jiangsu university, 2013.

[2] Zhang min ,Yu jian .Fuzzy clustering algorithm based on partition [J]. Journal of software, 2004, 15 (6) : 858-868.

[3] Yan ping Ran, Shao ping YU, Clustering algorithm based on hybrid model study [J]. Journal of henan science, 2012, 23 (3) : 324-324.

[4] Yue jia ,Shi tong Wang .Gaussian mixture model and the study of the EM algorithm and the initialization of the clustering [J]. Microcomputer information, 2006, (22) : 244-246.

[5] kai yuan Sheng .Clustering algorithm in the application research on large-scale data set [D]. Southern Yangtze university, 2014.

[6] Ling jie Zhang,Wei xiang XU.Association rule mining algorithm based on temporal constraint [J]. Computer engineering, 2012 (05) :50-52.