Mining on the subset of raw data set based on clustering

Yuling Ma[1,a]

[1]Information Engineering College, Shandong Yingcai college, Jinan, 250014, China

[a]myllz@sohu.com

Keywords: Big data era; Clustering algorithm; Association rule mining; ID3; Subspace; PAC learnable; Sample complexity.

Abstract. With the advancement of information process, the amount of the data accumulated by all walks of life is increasing exponentially. The emergence of massive data brings challenges to the traditional machine learning and data mining algorithms. In view of this problem, there have been many new researches, such as distributed machine learning, GPU acceleration processing, and the optimization of algorithms. But even so, when the amount of data is very big, for example, the data which come from biological field, mining on these data directly is still time-consuming and memory-consuming. In such big data era, what should we do first before mining? In this paper, we proposed mining subset method. It found out a representative subset of raw data through some related algorithms, and then applied data mining algorithms to the subset. Theory and experiments both verify the correctness of our method, especially when the dataset size is very large, the advantage of our method is more obvious.

Introduction

As we all known, data is indispensable to data mining. No data, No mine. In a sense, more data will get better mining results. But, if the amount of data is very large, training these data will be very time and memory consuming. Furthermore when the amount of data is increasing to certain extent, it perhaps generates data break point, after which mining on raw data is very low efficient. We believe that, if we mine the whole data set every time, it will be very time-consuming and be very difficult to tolerate especially when the amount of data is huge. If we can find a representative subset of redundant raw data, mining on the subset will obtain higher efficiency and similar mining results, so naturally it is a proper step to divide data before processing.

There are some researches about this, such as (Frey, B. J2007)[1]devised a method called "affinity propagation," which takes as input measures of similarity between pairs of data points. Real-valued messages are exchanged between data points until a high-quality set of exemplars and corresponding clusters gradually emerges. (Liu, G 2010)[2] proposed low-rank representation (LRR) to segment data drawn from a union of multiple linear subspaces. (Zhu et al,2015)[3] proposed an accelerated robust subset selection method, the robustness against outlier elements is greatly enhanced. And some other researches focus on feature subspace learning, which can be called dimension reduction. For example, (Qu et al,2005)[4] used association rule mining together with clustering method, by clustering features , divided raw dataset into some subset, and then mine on those subset ,so that it saved much time.

The rest of this paper is organized as follows: In section 1 we introduce the concept of clustering; in section 2 we propose the subset-mining method and give the theory analysis; in section 3, two kinds of algorithms are proposed by combining traditional machine learning algorithm and subset method; in section 4, we report on experiments; in section 5, we introduce some related work which follows by conclusion.

About clustering

Basic concepts. Clustering[5] is typical unsupervised learning paradigm. It divides data into several clusters through a certain metric standard. Thus similar data points are in the same cluster. Clustering method is wildly used in subspace learning, data visualization, dimension reduction and

outlier detection etc. Typical distance standards are cosine distance, Mahalanobis distance, Euclidean distance, Hamming distance and Binary distance etc. As an important learning paradigm, there are already some typical clustering algorithms, for example, k-means, k-medoids, Gaussian mixture model, spectrum clustering and so on.

k-means algorithm. Input：data set $X \subset R^n$ ， cluster number k . Initialize：randomly select k centers u1,…,uk . For i=1to k $C_i = \{x \mid i = \operatorname{argmin}_j \| x - u_j \|, \ x \in X\}$ $ui = \frac{1}{|C_i|}\sum_{x \in Ci} x$ Until hold some conditions(or ui is not change)；

Brief introduction of subset method

For some algorithms which need scan data set frequently, for example apriori algorithm, etc. When data size is very big, it will be very time-consuming and low efficient. If there are large amount of similar samples in raw data set, it is obviously a better choice that delete those redundant data before mining. Clustering algorithm is one of feasible methods which can efficiently eliminate redundant data. Especially after optimization, k-means algorithm can be executed high efficiently. So we propose the framework of mining subset. To those data sets that scale is very huge and have much similar sample, we could find a representative subset by clustering algorithm, and then mine on the subset.

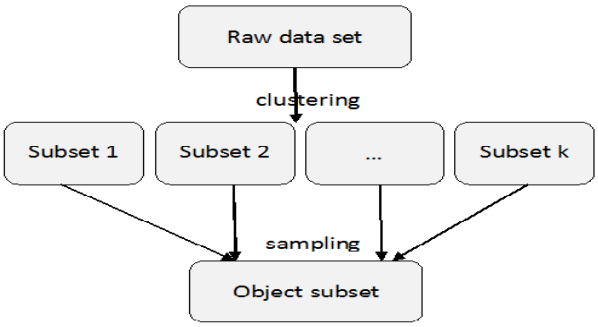

Fig. 1 frame work of subset method

Theory analysis. Now we introduce the definition sample complexity[6].

Corollary. Every finite hypothesis class is PAC learnable with sample complexity：mH(,δ) ≤ [log (|H|/δ)/ Ɛ].

Through sample complex corollary, we can see that if we have a certain number of samples (holds the condition $mH(,\delta) \leq [\log(|H|/\delta)/\text{Ɛ}]$), we can obtain valid mining result. But huge amount of data brings difficulty for computers, for example, if the dimension is very high ,it perhaps bring the problem of “dimension curse’. Furthermore , almost all algorithms have the time complexity O(f（N）)（N： the number of samples，f: a monotone increasing function about N）, so if N is extremely large, the consuming of time and memory is hard to tolerate for computers. It needs tradeoff between good mining result and time/memory complex. If we can find a strong expression （subset） of raw data， mining on the subset can get the similar result with the result of raw data.

Proposed methods

Subset-apriori algorithm. Since proposed by Agrawal et in 1993, association rule mining got the attention of researchers at home and abroad. Especially it is wildly used in basket analysis, disease association analysis, course system association research and so on. Thus association rule mining becomes one of most popular data mining algorithms. But its typical algorithm ， apriori，need scan data set frequently when being executed. It is very time consuming! Although some improved versions emerged later, for example, FP tree, method based on grid, method based on clique, etc, most of these versions are researched from the angle of data structure, and not think

about data set itself. So we apply the idea of subset mining to association rule mining. Experiments verify our method can improve mining efficiency significantly and obtain the similar result. The method can be extend to apply to other apriori algorithm versions.

Subset-ID3

Input：raw dataset D={(xi,yi)|i=1…N}, xi:instance，yi:label，cluster number: k，sample ratio: r

Output：tree

Procedure：

（1）Execute k-means on transaction data set D， obtain clusters $C_1,C_2,\ldots,C_k$

（2）for i=1：k

Sample from C_i according to the sample ratio r, get k sample subsets S_i。

（3）Obtain object subset S :

for i=1：k $S=\cup S_i$

（4）Mining on S through ID3 algorithm

Subset-apriori

Input：transaction data set T, Minimum support degree minsupp， minimum confidence degree minconf， cluster number k，sample ratio: r

Output: maximum frequency itemset L

Procedure：

（1）Execute k-means on transaction data set T，obtain clusters $C_1,C_2,\ldots,C_k$.

（2）for i=1：k

Sample from $C_i$according to the sample ratio r, get k sample subsets S_i。

（3）Obtain object subset S :for i=1：k $S=\cup S_i$

Subset –ID3. The same as above ,we apply subset method to decision tree classification. Attention that: we assumption that there are large amount of similar samples in data set. Decision tree learning [7]is one of the most widely used and practical methods for inductive inference. It is a method for approximating discrete-valued functions that is robust to noisy data and capable of learning disjunctive expressions. Now it mainly have three versions, for example,ID3, C4.5 and CART. Because every DT algorithm need large amount of computation when selecting the best attribute, it is very time-consuming. We propose Sub-ID3 algorithm, which first generate a representative subset of raw data, and then execute ID3 algorithm on it.

Experiments

We test proposed algorithms on computer of Intel(R) Core（TM）i5-4200M CPU@2.50GHz，4GB memory, Windows8.1，by using matalab. This paper experiment execute on a real dataset and benchmark dataset respectively.

Subset-apriori on a real dataset. With the support of a game corporation in Nan jing, we obtain their game tools transaction data which named GamePropsTransSet（GPTS）.The result is as Table 3:

Table 1 result of Subset-apriori algorithm on GPTS

data set GPTS						
Min-supp	Min-conf	sample ratio	Subset-apriori		run time（s）	
			recall	precision	apriori	s-apriori
0.2	0.5	0.1	1	1	214.56524	**126.2586**
0.3	0.5	0.2	1	1	15.887343	**7.463681**
0.4	0.6	0.15	0.875	1	4.409729	**1.787051**
0.5	0.8	0.01	1	1	0.696536	**0.372645**

From Table 1, we can see that the Subset-apriori algorithm can obtain good mining result, only need 10%~20% of raw data. Especially the threshold is big（minsupp=0.5，minconf=0.8），we only need one percent of raw data and obtain the same result with mining on raw data. Most important is that time and memory consuming is reduced significantly.

Subset-ID3 on benchmark data. We randomly select five benchmark data set from uci .The result is as follows:

Table 2 the result of Subset-ID3 algorithm on some benchmark data

dataset	size	ID3		Subset-ID3		
		Error rate	Time(s)	Error rate	Time(s)	Sample ratio
bank	4521*16	0.1309	0.507598	0.1175	**0.188631**	0.2
bankfull	45211*17	0.1349	6.841839	0.0524	**0.281168**	0.2
CTGS2	2126*36	0.0198	3.552113	0.0171	**0.255341**	0.6
CNAE-9	1080*856	0.1296	2.096919	0.1458	**1.450976**	0.6
Sensorless	58509*49	0.0156	91.95839	0.0198	**38.89846**	0.1

From table 2, we can see Subset-ID3 algorithm are less time-consuming than typical ID3. We highlight it in bold. Especially when the data size is big, for example , Sensorless data set, the amount of saved time is more obvious. And mining on the subset gets similar result.

Conclusion

Due to the explosive growth of data, subset selection methods are increasingly popular for a wide range of machine learning and computer vision applications .This kind of methods offer the potential to select a few highly representative samples or exemplars to describe the entire dataset. By only using the selected exemplars for succeeding tasks, the cost of memories and computational time will be greatly reduced. Furthermore we can use sample subset and feature subset simultaneously, and save more time and memory.

Acknowledgement

The work is supported by three funds as follows: University Science and technology project of Shandong province（J15LN55）, Vocational education and adult education project of Shandong province (2014zcj015) and the education reform project of Shandong province (YCXY-X2014011).

References

[1] Frey, B.J, Dueck.D: Science 315(5814) (2007), p.972–976

[2] Liu, G Lin, Z. Yu: In ICML(2010) , p. 663–670

[3] Feiyun Zhu, Bin Fan, Xinliang Zhu:AAAI (2015), p. 3660

[4] Shouning Qu，Caiyun Dong，Dejun Xu，Tong Wu：Computer system and application(2005)，Vol.14(4)，p.20-22

[5] Jiawei Han:<Data mining-concept and technology>2th version(2006)，p. 383-402

[6] Shai Shalev-Shwartz, Shai Ben-David: < Understanding machine learning > , (2014), p. 22-23

[7] Tom M. Mitchell,<Machine Learning> (March 1, 1997), p.52-54