Research Paper Presentation

Diya Goyal

CS20BTECH11014

Title

Fast Cluster-Learning with prior probability from Big Dataset

Authors

- Tengyue Li
- Simon Fong
- Joao Alexandre Lobo Marques
- Raymond K Wong

Terms to know

- **Apriori Algorithm :** Apriori is an algorithm for frequent item set mining and association rule learning over relational databases.
- **Association Rule Mining :** It is the data mining process of finding the rules that may govern associations and causal objects between sets of items.
- **Clustering :** In machine learning, we often group examples as a first step to understand a subject (data set) in a machine learning system. Grouping unlabeled examples is called clustering.

Terms to know

- **Preprocessing :** Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.
- **Prior Probability :** A probability as assessed before making reference to certain relevant observations, especially subjectively or on the assumption that all possible outcomes be given the same probability.

Introduction

- The association rule mining algorithms, such as Apriori method for instance and other successors were invented in the late 90's. They were not designed to work with big data in mind in those days.
- Apriori algorithm, being a classical association rule mining algorithm is known to have a drawback in scalability. The time performance in big O notation would have it scale up exponentially to the dimensions or the number of attributes of the data.
- In order to tackle the challenges of association rule mining over big data or big dataset where the data are usually presented as a two-dimensional data matrix, some modifications are needed either at the algorithmic level or at the data processing level.

- Earlier on, the authors advocated an alternative preprocessing method which aims at improving the quality of association rules.
- It first segments the full dataset into clusters prior to running the association rule mining algorithm on. The underlying concept is to mine the association rules from subsets of the data which are supposedly more concentrated in data that occur frequently together. This is because clustering helps pulling data that are similar to each other together, reducing the distance space between the scattered data throughout the whole dataset.
- By this simple idea, the overall association rules quality improves. But the problem is, in advance there are no way to know which is most appropriate to be mined by association rule mining algorithm.
- Although the cluster size is smaller than the original dataset size, associate rule mining algorithm such as Apriori still takes time.

Proposed Methodology

- Traditional association rule mining loads the whole dataset for generating association rules.
- To get around this time-consuming process, the rule generated process is controlled by harvesting only a certain number of rules which usually are the top ones.
- Alternatively, a subset of dataset is used for the association rule mining instead of the whole one. In this case, out of several clusters or subsets of the original dataset that was partitioned by some clustering algorithm, there should be one or few out of all clusters should be used for incomplete association rule mining.
- The qualities of the clusters vary as there is no rule-of-thumb exists in either assuring the clusters are made all suitable for association rule mining or knowing in advance which cluster is most suitable for highest quality rules to be mined.

- As a solution, the proposed process advocates using a statistical property called Prior Probability to be used. Hence it is no longer necessary to try out every cluster, because a user can base on the Prior Probability (PP) to decide which cluster or a short-listed set of clusters to be selected for association rule mining. Thus speeding up the process.
- Due to the speed saving, this association rule mining strategy is called Fast Cluster Learning (FCL).
- FCL divides up the large dataset into small fragments, not randomly, but according to the similar data which shall be grouped together and PP is calculated for each of these clusters.
- With the PP value in place for each cluster, the cluster candidates can be sorted in descending order. The cluster candidate that has the highest PP value is suggested to be chosen for immediate association rule mining.

- In the methodology, parameter free clustering algorithm namely Expectation Maximization (EM) algorithm is suggested to be used. It will automatically find the optimal number of clusters by its maximization mechanism.
- For brevity, the data items only have a single numeric attribute whose values distributed Normally over all the k clusters, $C_1, C_2..C_j..C_k$ where $j\epsilon[1, k]$. In a very primitive state of probabilistic clustering, the following are the parameters of the model that need to be calculated:

$$\text{Model\_Param} = \text{Model}(\text{pp}_j, \mu_j, \sigma_j, \text{where } j\epsilon[1, k])$$

- At the beginning the Model_Param values are set randomly; based on the given the current parameter values, the probability of the cluster membership for each data item is computed; the Model_Param values are re-evaluated and updated using the calculated probabilities which are shown to have yielded better cluster memberships. This iterative process repeats until the Model_Param appears to converge where no further improvement on the cluster membership can be observed.

Key Phases

Phase 1 : Expectation step

- The probability of each data item x_i where $i \epsilon [1, n]$ from the dataset is computed for finding out the best cluster membership for the cluster c_j where $j \epsilon [1, k]$ using

$\Pr(E) \forall_{i,j} = \varepsilon_{i,j} = pp_j \times \Pr(x_i | c_j)$
where the cluster pp_j is computed in each j^{th} iteration based on the model parameters values and $\Pr(x_i | c_j)$ is calculated from the Normal distribution of the j^{th} cluster where $j \epsilon [1, k]$ given that the current model is configured with the $Model\_Param \epsilon (x_j, \mu_j, \sigma_j)$

Phase 2 : Maximization step

- Calculates the most suitable Model_Param values in order to maximize the likelihood of models that are guessed to match the current positions or memberships of the data items and the model parameters are re-evaluated according to

Prior probability: $pp_{i,j} = \sum_i \frac{\varepsilon_{i,j}}{n}$

Mean: $\mu_{i,j} = \frac{\sum_i x_i \times \varepsilon_{i,j}}{\sum_i \varepsilon_{i,j}}$

Standard deviation: $\sigma_{i,j} = \sqrt{\frac{\sum_i x_i - \mu_j{}^2 \times \varepsilon_{i,j}}{\sum_i \varepsilon_{i,j}}}$

- These dual steps repeatedly increase the log-likelihood of all the clusters until there is no more significant refinement according to :

$$\log \Pr(x) = \log \sum_j (\Pr(x|c_j) \times pp_j)$$

- Usually the overall quality of the clusters which is represented by log-likelihood will rise sharply at the beginning over some initial iterations.
- In our fast cluster learning methodology, EM(Expectation Maximization) is chosen because it is guaranteed to converge to highest possible log-likelihood fitness.

Limitation

- The only limitation is the time consumption which can be significantly large. It is because there exist local maximum and global maximum.
- For prevention of falling into local maximum, the EM algorithm is programmed to run several times for obtaining some chances of reaching the global maximum as each time they start with different random orientation for the initial clusters and model parameters, guessing what their suitable parameter values are.
- Hence there is a compromise e between using a heavy algorithm that takes considerable amount of time to run and achieving the best clusters most of the time without specifying the k parameter value.

Conclusions

- In this paper, a new data mining methodology called Fast Cluster Learning is proposed.
- It is designed for enhancing the quality of rules by association rule mining such as Apriori method.
- In this paper, a simple quality indicator called Prior Probability (PP) is proposed to use for quickly identifying a cluster that would be useful for subsequent rule mining.