Algorithms and Optimization for Big Data

ENDSEM 2017

Hardil Mehta(1401018)
Information and Communication technology (ICT)
School of Engineering and Applied Science
Ahmedabad,
Email: Hardil.m.btechi14@ahduni.edu.in

***Abstract*— In order to recommend relevant content to a user, many web companies use logistic regression models to predict the probability of the user's interest in an item. In scenarios where the data is abundant, having a more fine-grained model at the user or item level would potentially lead to more accurate prediction, as the user's personal preferences on items and the item's specific attraction for users can be better captured. Our work aims to find out relationships between jobs and people skills making use of data from LinkedIn users' public profiles. Recommender systems are systems that, based on information about a user's past patterns and consumption patterns in general, recommend new items to the user. Some systems incorporate information about the items in question, others are based only on usage patterns; the latter kind of system is known as a collaborative filtering system. Instead of asking the user to explicitly pick filters for a search, collaborative filtering uses information about the user's past behavior and similar users to make suggestions.**

Keywords—Recommender system; Generalized linear model; ParallelBlock-wiseCoordinateDescent; association rules mining; collaborative filtering;

I. INTRODUCTION

The proliferation of data and information-rich user experiences have transformed data mining into a core production use case, especially in the consumer web space. A typical example is showcasing relationships between pairs of items based on the wisdom of the crowd, also known as item-to-item collaborative filtering (ICF) [6]. At LinkedIn, the largest online professional social network, item-to-item collaborative filtering is used for people, job, company, group, and other entity recommendations and is a principal component of engagement. That is, for each entity type on the site, there exists a navigational aid that allows members to browse and discover other content. Our aim was to build a collaborative filtering system to recommend skill for User at LinkedIn. The end product would allow a current user to enter his or her transcript and - based on which skills he had Acquired and what career goal he has - a list of Skills and career in which the User would potentially do well would be returned. There are three important criteria that determine how useful an algorithm is.

A. Primary Criteria

(i) Quality of Prediction
(ii) Speed/Scalability
(iii) Online Updating

B. Secondary Criteria

Sparse data handling - sometimes our datasets are very sparse, and but we still want to make good predictions.

II. GENERALIZED LINEAR MODEL ALGORITHM

Generalized linear model (GLM) is a widely used class of models for statistical inference and response prediction problems. One common approach is to introduce ID-level regression coefficients in addition to the global regression coefficients in a GLM setting, and such models are called generalized linear mixed models (GLMix) in the statistical literature. However, for big data sets with a large number of ID-level coefficients, fitting a GLMix model can be computationally challenging. Here an approach is taken use from [7] which successfully overcame the scalability bottleneck by applying parallelized block coordinate descent under the Bulk Synchronous Parallel (BSP) paradigm.

1) GLMixModel

Now we consider the GLMix model for the career path recommendation problem. To measure whether career j is a good match for a member m and to select the best career. Let y_{mjt} denote the binary response of whether member m would apply for career j in context t, where the context usually includes the skills and experience of the user profile. We use q_m to denote the feature vector of member m, which includes the features extracted from the member's public profile, e.g., the member's title, job function, education history, industry, etc. We use s_j to denote the feature vector of career j, which includes features extracted from the career requirements, e.g. the job title, desired skills and experiences, etc. Let x_{mjt} represent the overall feature vector for the (m,j,t) triple, which can include q_m and s_j for feature-level main effects, the outer product between q_m and s_j for interactions among member and job features, and features of the context. The GLMix model for predicting the probability of member m applying for job j using logistic regression is:

$$g(E[y_{mjt}]) = x^T b + s^T a_m + q^T \beta_j$$

Where $g(E[y_{mjt}]) = \log(E[y_{mjt}]/1- E[y_{mjt}])$ is the link function, b is the global coefficient vector (also called fixed effect coefficients in the statistical literature); and a_m and βj are the coefficient vectors specific to member m and career j.

2) Priors: To mitigate the risk of overfitting due to the large number of parameters and sparsity of the data, we put the following Gaussian priors on both fixed effects and random effects: b ~ N(0, 1 λb I), αm ~ N(0, 1 λα I), βj ~ N(0, 1 λβ I), (2) where I is the identity matrix, 1 λb is the prior variance of b, and 1 λα and 1 λβ are the prior variances of αm and βj.

3) General Formulation

We now consider a more generic formulation of GLMix where more than two sets of random effects can be present in the model. This is useful for scenarios such as multi-context modeling, where we can have random effects to model the interactions between the context id and user/item features. The general formulation of GLMix is provided in the following equation:

$$g(E[y_n]) = \boldsymbol{x}_n'\boldsymbol{b} + \sum_{r\in\mathcal{R}} \boldsymbol{z}_{rn}'\boldsymbol{\gamma}_{r,i(r,n)},$$
$$\boldsymbol{b} \sim p(\boldsymbol{b}), \quad \boldsymbol{\gamma}_{rl} \sim p(\boldsymbol{\gamma}_{rl}), \quad \forall r \in \mathcal{R}, 1 \le l \le N_r$$

We use γr,i(r,n) to denote random effect coefficient vector and z_{rn} as the corresponding feature vector, for random effect type r in the n-th sample.

$$s_n = \boldsymbol{x}_n'\boldsymbol{b} + \sum_{r\in\mathcal{R}} \boldsymbol{z}_{rn}'\boldsymbol{\gamma}_{r,i(r,n)},$$

In Algorithm 1, a parallel block-wise coordinate descent based iterative conditional mode algorithm is proposed, where the posterior mode of the random effect coefficients Γr for each random effect r is treated as a block-wise coordinate to be optimized in the space of unknown parameters. Given the scores s defined above, the optimization problems for updating the fixed effects b and the random effects Γ are provided in Equation below:

$$\boldsymbol{b} = \arg\max_{\boldsymbol{b}} \left\{ \log p(\boldsymbol{b}) + \sum_{n\in\Omega} \log p(y_n | s_n - \boldsymbol{x}_n'\boldsymbol{b}^{old} + \boldsymbol{x}_n'\boldsymbol{b}) \right\}$$
$$\boldsymbol{\gamma}_{rl} = \arg\max_{\boldsymbol{\gamma}_{rl}} \left\{ \log p(\boldsymbol{\gamma}_{rl}) + \sum_{n|i(r,n)=l} \log p(y_n | s_n - \boldsymbol{z}_{rn}'\boldsymbol{\gamma}_{rl}^{old} + \boldsymbol{z}_{rn}'\boldsymbol{\gamma}_{rl}) \right\}$$

Algorithm 1: Parallel block-wise coordinate descent (PBCD) for GLMix

```
while not converged do
    Update fixed effect b as in (7)
    foreach n ∈ Ω in parallel do
        Update s_n as in (9)
    foreach r ∈ R do
        foreach l ∈ {1, ..., N_r} in parallel do
            Update random effect γ_rl as in (8)
        foreach n ∈ Ω in parallel do
            Update s_n as in (10)
```

III. Collabarative Filtering Using Associative Rules

A. Fundamentals of Association Rules

Association rules try to connect the causal relationships between items. An association rule essentially is of the form A1, A2, A3,…. => B1, B2, B3, ... It attempts to show how a series of items can determine another series of items. For a more concrete example, if we said A => B, C, that would mean that the appearance of item A in someone's history would imply that B and C would be there as well.

It's not just the items that matter, however; another important factor is the confidence of a rule. Confidence is the intuitive idea of how applicable a rule is. It can range from 0 to 1. If the confidence is 1, then we know that the rule always applies - that is, every time we see A, we also see B and C. However, if the confidence is 0, it means it's never correct - A does not imply B and C.

B. Association Rules for Recommendations

For our purposes we used association rules of the form A => D. This means that we looked at all skill relationships. That is, what is the likelihood of the user requiring a skill D, given that the active user has rated A?

We create a square matrix of all these single-item relationships and their associated confidence values between all n items in the dataset. Then, we treat the user as a vector in n-dimensional space. If you multiply the matrix by the vector, you get what is called a recommendation vector - the most likely skill that the user will require, given the skill they already have.

You can easily use is recommendation vector to order preferences of a user.

C. Advantages

- It is incredibly fast. Building the matrix takes a very short amount of time, and then all recommendations after that are instantaneous.
- It generates credible results - what you might think of recommending a person given what skills they have.
- It works well with sparse data sets, especially if we implement a multi-level association rule index which has higher levels of generalization (in case the lower ones don't have enough information).

D. User-skill table

User-skill Table

Skill / User	I1	I2	… …	In
U1	r11	r12	… …	r1n
U2	r21	r22	… …	r2n
… …	… …	… …	… …	… …
Um	rm1	rm2	… …	rmn

E. Measuring the skill Similarity and Selecting neighbours.

There are several similarities in the algorithm Collaborative filtering algorithms. Pearson correlation, cosine vector similarity, adjusted cosine vector similarity, mean square deviation and Spearman correlation. Pearson's correlation, as following formula, measures the linear correlation between two vectors of skills.

Choose a neighbor who will serve as a referee. Both techniques have been employed in the collaborative filtering recommendation system. Threshold-based selection,

According to the the similarity of the user exceeds a critical value as the target user's neighbors think.

Top - N technology, the best nitrogen neighbors and N is given first.

F. Association rules mining

The apriori is the important algorithm in the algorithms of association rules mining. The main idea of the apriori is scanning the database repeatedly. The most important step in mining association is the generation of frequent item sets. In apriori algorithm, most time is consumed for scanning the database repeatedly.

Let I = { i1 , i2 , …, im } be a set of all items, where an item is an object with some predefined attributes. A transaction T = < tid , It > is a duple, where tid is the identifier of the transaction. A transaction database T consist s of a set of transactions. An itemset is a subset of the set of items.

Definition 1: An association rule takes the form X =>Y where X < I , Y < I , and X ∩Y = O. The support of the rule X =>Y in the transaction database is :

support (X =>Y) = | { T :X ∪ Y ∪ T , T ∈ D} | / | D |

Definition 2: The confidence of the rule X =>Y in transaction database is :

confidence (X => Y) = | { T :X ∪ Y ∪ T , T ∈ D} | / | { T :X < T , T ∈ D} |.

G. Producing Prediction

Collaborative filtering user-skill data are usually represented as preference matrixes. They will change the transaction database for mining association rules. Each transaction includes a transaction ID and content. TID is the transaction ID of the user's user ID for the transaction to which they belong. The content of the item ID and assessments have been evaluated by the user.

The Apriori algorithm calculates the frequent item sets in a database using many repeated iterations. All the frequent item sets calculated in the ith iteration are called k skill sets. Each iteration consists of two steps: generating the candidate skill sets, and calculating and choosing the candidate skill sets. Its kernel thought is as follows

(1) L1 = {Large 1-Skill };
(2) for (k = 2; kk - 1 ! = 0; k + +)
(3) Ck = Apriori-gen (Lk - 1)
(4) for all transaction t∈D do begin
(5) Ct = SubSet (Ck , t) ;
(6) for all candidates c∈Ct do
(7) c. count + +;
(8) end
(9) Lk = { c∈Ck };
(10) end
(11) UkLk
(12) end

IV. RESULTS

```
>> Module1
--------------------MODULE 1---------------------
             Career path suggestion
-------------------------------------------------
Software Engineer
Automation tester
UI Developer
>>
```

Fig 1: Module 1 suggesting a career path

```
--------------------MODULE 2---------------------
              skillset suggestion
-------------------------------------------------
Html
css5
Ruby on rails
```

Fig 2: Module 2 Suggesting skillsets

V. CONCLUSION

Two approach for solving the problem is discussed above. Generalized linear models (GLMs) have been successfully applied to a wide range of applications such as response prediction and collaborative filtering recommendation method Combining the association rules mining can be used for recommendation system.

REFERENCES

[1] J.S. Breese, D.Heckerman, and C.Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the Fourteenth Conference on Uncertainty in Artifical Intelligence, 1998.

[2] M.Deshpande and G. Karypis. Item-based top-n recommendation algorithms. ACM Trans. Inf. Syst., 22(1):143-177, 2004.

[3] B.M. Sarwar, G. Karypis, J.A. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th International World Wide Web Conference, pages 285-295, 2001.

[4] C. Kim and J. Kim. A recommendation algorithm using multi-level association rules. In Proceedings of Web Intelligence 2003., 2003.

[5] K.AliandW.vanStam. Tivo:Makingshowrecommendations using a distributed collaborative filtering architecture. KDD, pages 394–401, 2004.

[6] X. Su and T. M. Khoshgoftaar. A survey of collaborative filtering techniques. Advances in Artificial Intelligence, 2009, Jan. 2009.

[7] Xianxing Zhang, Yitong Zhou, Yiming Ma, Bee-Chung Chen, Liang Zhang, Deepak Agarwal. GLMix: Generalized Linear Mixed Models For Large-Scale Response Prediction

[8] D.Pennock, E.Horvitz, S.Lawrence, and C.L. Giles. Collaborative filtering by personality diagnosis: A hybrid memory- and model-based approach. In Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, UAI 2000, pages 473-480, Standford, CA, 2000.