TELKOMNIKA, Vol.11, No.3, March 2013, pp. 1213 ~ 1220
ISSN: 2302-4046

Information Extraction from Research Papers based on Conditional Random Field Model

Zhu Shuxin*, Xie Zhonghong, Chen Yuehong
College of Information Science and Technology, Nanjing Agricultural University
Weigang 1, Nanjing, China, 210095
*Corresponding author, e-mail: zsx@njau.edu.cn

Abstract

With the increasing use of CiteSeer academic search engines, the accuracy of such systems has become more and more important. The paper adopts the improved particle swarm optimization algorithm for training conditional random field model and applies it into the research papers' title and citation retrieval. The improved particl swarm optimization algorithm brings the particle swarm aggregation to prevent particle swarm from being plunged into local convergence too early, and uses the linear inertia factor and learning factor to update particle rate. It can control algorithm in infinite iteration by the iteration between particle relative position change rate. The results of which using the standard research papers' heads and references to evaluate the trained conditional random field model shows that compared with traditionally conditional random field model and Hidden Markov Model, the conditional random field model, optimized and trained by improved particle swarm, has been better ameliorated in the aspect of F1 mean error and word error rate.

Key Words: *Conditional Random Fields Model; Particle Swarm Optimization; Maximum Likelihood Parameter Estimation; Information Extraction*

1. Introduction

Academic search engines such as CiteSeer and Cora has brought great convenience and influence for researchers. So the quality provided by such systems is very important. It is dependant on the information retrieval components which extract titles, authors, institutions and other meta-data from papers because these metadata later would be used for a variety of applications, such as domain based searching and the authors research.

Previously, academic information retrieval is based on two main machine learning techniques. The first is Hidden Markov Model (HMM). HMM obtains generation model from the input and label sequence pairs. Although HMM gains great success, the standard HMM model is difficult to build the model for multiple independent features of observation sequence. The other technique is based on support vector machine (SVM) classifier. It can process multiple non-independent features, but it divides information retrieval into two steps and breaks the close coordination between state transition and observation.

This paper introduces a conditional random field model [1] (CRF) for the academic metadata retrieval and applies the model to solving the practical problems in order to prove that the model is superior to HMM and SVM. CRF has been successfully used in biological entity recognition [2, 3, 4]. The model loosens the strong random hypothesis of HMM and effectively overcomes the label bias problem. Similar to Max Entropy Markov Model (MEMM). CRF is also a conditional probability sequence model. But the difference is that CRF is an undirected graph model.

In CRF, maximum likelihood parameter estimation is crucial because these parameters concern the performance and accuracy of the applications based on CRF. The maximum likelihood estimation usually adopts nonlinear conjugate gradient algorithm [5, 6, 7], Newton method [8], BFGS [9], gradient acceleration algorithm [11], virtual evidence acceleration algorithm [1], piecewise hypothesis likelihood method [10, 11], stochastic gradient method [12], minimum divergence beam method [13] and so on.

This paper uses improved particle swarm optimization algorithm to estimate the maximum likelihood parameters of CRF model. In order to prevent the algorithm fallen into the

Received October 5, 2012; Revised January 5, 2013; Accepted January 18, 2013

local convergence in the early time of search, this paper adopts an appropriate aggregation degree to control particles aggregation level. And to prevent the algorithm converging slowly near the best location and going into an infinite iteration, this paper employs inertia and learning factor which is linear varying to control particle search range and the relative change rate of iterative particle position to terminate iteration. Finally, this article applies the trained conditional random field model is the use of the standard research papers' head and reference data set retrieval. The experimental results shows that compared with traditionally conditional random field model and Hidden Markov Model, the conditional random field model ,optimized and trained by improved particle swarm, has been better ameliorated in the aspect of word accuracy and F metrics.

This paper is organized as follows: Section 1 introduces the concept of CRF and its' maximum likelihood parameter estimation, Section 2 describes the improved particle swarm optimization and proposes a new model parameters estimation algorithm of CRF based on it, Section 3 shows the experimental results and Section 4 gives the summarization of this paper.

2. Conditional Random Field Model

Conditional random field model (CRF) is an undirected graph model used to calculated the conditional probability of the output sequence under the premise of the given input sequence. Lafferty [1] gives the definition of CRF: For a given observation sequence x, the conditional probability of its' tag sequence y is the normalized product of the bit function, as shown in Formula (1).

$$p(y \mid x) \propto \exp(\sum_j \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_k \mu_k s_k(y_i, x, i)) \quad (1)$$

where $t_j(y_{i-1}, y_i, x, i)$ is the transition feature function of the whole observation sequence, y_i is the label of the tag sequence I, $s_k(y_i, x, i)$ is the state feature function of i , x is the observation sequence， λ_j and μ_k are respectively the weight parameter related to the transition feature function and state feature function.

Common undirected graph model is a linear chain and corresponds to a finite state machine which is suitable for tag sequence. We simply the transition feature function and state feature function as below:

$$s(y_i, x, i) = s(y_{i-1}, y_i, x, i)$$
$$F_j(y, x) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, x, i) \quad (2)$$

where $f_j(y_{i-1}, y_i, x, i)$ could be taken as state function $s(y_{i-1}, y_i, x, i)$ or transition function $t_j(y_{i-1}, y_i, x, i)$ .Thus, for an observation sequence X, the probability of its' tag sequence y is expressed as follows:

$$p(y \mid x, \lambda) = \frac{1}{Z(x)} \exp(\sum_j \lambda_j F_j(y, x)) \quad (3)$$

Among them, Z (X) is a normalization factor:

$$Z(x) = \sum_{x,y} \exp(\sum_k \lambda_k f_k(c, y_c, x)) \quad (4)$$

For the maximum likelihood parameter estimation of the conditional random field model as shown in Formula(2), it is just a process by training dataset $D = \{(y^{(1)}, x^{(1)}), ..., (y^{(n)}, x^{(n)})\}$ to estimate the parameter $\lambda = (\lambda_1, \lambda_2, ...)$. Among them, D is generated by empirical distribution $\tilde{p}(x, y)$ in order to maximize the log likelihood of the training data.

3. Conditional Random Field Model Parameter Estimation based on Improved Particle Swarm Optimization

3.1. Improved Particle Swarm Optimization Algorithm

Particle swarm optimization algorithm [14], first designed by Kennedy and Eberhart, imitates birds' prey behavior to solve the optimization problem. Particle swarm optimization algorithm makes a particle swarm (x_i) search the optimal value according to a certain evolution rule by initializing it. In each round of evolution, each particle's velocity and position have been updated in accordance with their current velocity and position, local and global optima value, shown as follows:

$$\begin{aligned} &v_{id}(t+1) = w * v_{id}(t) + \\ &c_1 * rand() * (p_{id} - x_{id}) + \\ &c_2 * rand() * (p_g - x_{id}) \end{aligned} \tag{5}$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \tag{6}$$

Among them， p_{id} ,the local optimal value, refers to the optimal value got by particle i in the dimension d hitherto. p_g ,the global optimal value，refers to the optimal value got by all particles searching the whole search space up to now. $v_{id}(t)$ stands for particle i's velocity in dimension d and $x_{id}(t)$ represents particle i's current location in dimension d. $x_{id}(t+1)$ denotes particle i's next location in dimension d. w is the inertia factor and c_1 is the local learning factor. c_2 is the global learning factor and $rand()$ is the random function in the range of [0,1].

The traditional particle swarm optimization algorithm has the following problems:

1. Premature convergence to the local optimal value instead of the global

In order to prevent the search prematurely into local convergence, we use particle swarm aggregation [15] to control it. Particle swarm aggregation describes particle dispersion defined as:

$$d(t) = \max\{| x_{id} - x_{jd} |, i, j = 1, 2, ..., m; i \neq j; d = 1, 2, ..., N\}$$

Among them，m stands for the size of the particle swarm and n represents the search space dimension. x_{id} is denotes the particle i in dimension d and x_{jd} denotes the particle j. If $d(t)$ is less than the set threshold value e, then we reinitialize particle swarm's velocity and position in d-dimensional space.

2. Particle swarm searches initially very quickly, however, when close to the optimum position, its searching speed is slow, and even in the risk of being trapped into local unlimited iteration.

In the particle swarm optimization algorithm, parameter w, c_1, c_2 is crucial to the speed of search and convergence. A big inertia factor does good to global search and a small inertia factor is beneficial to local search. Its' experiential value is close to 1 in the early search stage and close to 0.4 in the final stage. This paper chooses the linearly decreasing inertia factor in

order to acquire better global search ability in the early stage and better local search ability in the final stage.

$$W(t) = 1 - 0.6*t/M \quad (t = 0 \sim M - 1) \tag{7}$$

Where M is the maximum number of iterations and t is the current iteration.

Traditional particle swarm optimization algorithms usually set c_1 and c_2 as constant number. In order to make the particle swarm converge as soon as possible in the early large search space without explosive proliferation and converge to the optimal value as soon as possible in the final small search space without revolving around the optimal value, we select Formula (8) to update c_1 and c_2:

$$c_1 = 2 - \frac{t}{M}, c_2 = 2 + \frac{t}{M} \tag{8}$$

Finally, to ensure that the algorithm is not trapped into infinite iteration near the optimal position, the paper adopts the rate of relative position change to terminate iteration. When D is less than the threshold value e, iteration will be canceled.

$$D = (x_{id}(t+1) - x_{id}(t)) - (x_{id}(t) - xid(t-1)) / (x_{id}(t) - xid(t-1)) \tag{9}$$

Among the, $x_{id}(t)$ stands for the current position of particle i in dimension d and $x_{id}(t+1)$ represents the next position of particle i in dimension d and xid(t-1) is the previous position.

3.2. Conditional Random Fields Model Parameter Estimation Method Based on the Improved Particle Swarm Optimization Algorithm

The paper lets the parameter vector λ of the conditional random field model of as particle group and make them in a d - dimensional space search for optimal values for training conditional random fieldsmodel. Detailed steps are described as follows:

First step: Supposing the current iteration $t = 0, v_{id}(0) = 0$ and the dimension number of search space is d, we initialize the particle swarm X containing m particles:

$$x_{id} = \{\lambda_{1d}, \lambda_{2d}, ..., \lambda_{nd}\}$$

$$= \{random_1, random_2, ..., random_n\}$$

$$(i = 0 \sim m - 1)$$

We set the local optimal position and global optimal position as: $p_{id} = x_{id}, p_g = 0$. Then we preserve $fitness_{id}$,the local adaptive optimal value, in the array $F[i][d]$ and save the global adaptive optimal value in the variable F_g.

Second step: Updating the local and global optimal position

Calculating the adaptive value of each particle:

$$fitness_{id} = \tilde{p}(x, y)\sum_j \lambda_j F_j(y,x) - \tilde{p}(x)\log Z(x)$$

We compare the adaptive value of particle I in dimension d with the value of array $F[i][d]$.If the current value is greater than $F[i][d]$, then we reset $F[i][d]$ as the current value. If the current value is greater than F_g, then we reset F_g as the current value.

Third-step: updating the search velocity and position of the particle swarm:

Vid(t+1) = (1- 0.6*t/M) * Vid(t) + (2-t/M)*rand()*(Pid - Xid) + (2+t/M)*rand()*(Pg - Xid)

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1)$$

Forth-setp：calculating the particle swarm's aggregation $d(t)$:

$$d(t) = \max\{| x_{id} - x_{jd} |, i, j = 1, 2, ..., m; i \neq j\}$$

$$=\max\{\sqrt{(\lambda_{di1}-\lambda_{dj1})^2+...+(\lambda_{din}-\lambda_{djn})^2}, i, j=1,2,...,m; i\neq j\}$$

$$= \max\{\sqrt{\sum_{k=1}^{n}(\lambda_{dik}-\lambda_{djk})^2}, i, j = 1, 2, ..., m; i \neq j\}$$

If $d(t)$ is less than the preset threshold value e, then we reinitialize the particle swarm's velocity and position in the d-dimensional space.

Fifth-step: $t = t + 1$.

If D< 10-7 or t = M, then the iteration will be terminated. Otherwise turn to the second-step.

4. Experiment

4.1. Experimental Data

We selects two research papers' data sets as our experimental data. One is containing the head portion and the other is containing citations from references. The two data sets have been used for multiple research standard test data（McCallum et al., 2000; Han et al., 2003.

4.1.1. Data Set of Research Paper's Header

Academic thesis title is defined as all words from the start of the paper to the first chapter (usually the introduction) or to the end of the first page. The title section contains 15 retrieval domain: title, author, unit, address, summary, e-mail, date, abstract, introduction, telephone number, keyword, URL, degree, ISSN, page number. The header data set contains 935 headers. According to the previous study, we randomly select 500 pieces of data as training data set and the remaining 435 pieces as test data set. We call the data set H.

4.1.2. Data Set of Research Paper's References

References data set is generated by Cora (McCallum et al., 2000). It contains 500 reference items and we selects 350 items of them as the training data set. The remaining 150 items are treated as test data. References contain 13 domains: author, title, editor, book title, date, journal, volume, school, institutions, page number, address, press, summary. We call the data set R.

4.2. Validation Technique

In order to get a comprehensive evaluation, we used different methods to compare the test results. In addition to adopting previously used word accuracy measurement method, we will also use a domain F measurement method. The two methods help each other and do better to the evaluation results.

1. Word accuracy. The number of the words is defined as A which belong to a particular domain and are correctly identified in the domain. The number of the words is defined as B which not belong to but identified in a particular domain. The number of the words is defined as C which belong to a particular domain but are not correctly identified in the domain and is D which not belong to and are not identified in a domain. The word accuracy is calculated by $\frac{A+D}{A+B+C+D}$. The ratio of the number of the words whose redictive identification and real identification are the same to the total number of words, is defined as the comprehensive accuracy.

2. F1measurement. We use the precision rate and recall rate to calculate F1 measurement. Precision rate is expressed as $\frac{A}{A+C}$ and recall rate is denoted as $\frac{A}{A+B}$.F1 is represented as $\frac{2\times Precision\times Recall}{Precision+Recall}$. All domains' average F1 measurement is defined as the average F metric.

4.3. Experiment Result

This paper lists the comparison results among the conditional random field model trained by the improved particle swarm optimization algorithm, HMM and the conditional random field mode trained by the traditional particle swarm optimization algorithm. Table 1 reports the results based on H data set. Similar to the previous experiments（Seymore et al., 1999; Han et al., 2003, it excludes the introduction field and page number field. F1 measurement results is calculated by the precision rate and recall rate of the original experimental report. Table2 reports the results based on R data set. CRF（P） denotes the conditional random field model trained by the improved particle swarm optimization algorithm and CRF（T） denotes the traditional conditional random field model. It could be found from the comparison of data , the conditional random field model trained by the improved particle swarm optimization algorithm behaves as good as or better than HMM and the traditional conditional random field model in the aspect of word accuracy, F1 measurement. Moreover it is significantly better than the other two models in F measurement and comprehensive accuracy.

Table 1. Test results based on H data set

	HMM		CRF（P）		CRF（T）	
	acc.	F1	acc.	F1	acc.	F1
title	98.2	82.2	99.7	97.1	98.9	96.5
author	98.7	81.0	99.8	97.5	99.3	97.2
unit	98.3	85.1	99.7	**97.0**	98.1	93.8
address	99.1	84.8	99.7	**95.8**	99.1	94.7
summary	97.8	81.4	98.8	**91.2**	95.5	81.6
e-mail	99.9	92.5	99.9	**95.3**	99.6	91.7
date	99.8	80.6	99.9	**95.0**	99.7	90.2
abstract	97.1	98.0	99.6	**99.7**	97.5	93.8
Telephone-number	99.8	53.8	99.9	**97.9**	99.9	92.4
keyword	98.7	40.6	99.7	**88.8**	99.2	88.5
URL	99.9	68.6	99.9	**94.1**	99.9	92.4
degree	99.5	68.8	99.8	**84.9**	99.5	70.1
ISSN	99.8	64.2	99.9	86.6	99.9	**89.2**
Average F		75.6		**93.9**		89.7
Comprehensive accuracy	93.1%		**98.3**%		92.9%	

Table 2. Test results based on R data set

	HMM		CRF（P）		CRF（T）	
	acc.	F1	acc.	F1	acc.	F1
author	96.8	92.7	99.9	**99.4**	98.9	96.5
Book title	94.4	0.85	97.7	**93.7**	99.3	97.2
date	99.7	96.9	99.8	**98.9**	98.1	93.8
editor	98.8	70.8	99.5	**87.7**	99.1	94.7
institution	98.5	72.3	99.7	**94.0**	95.5	81.6
journal	96.6	67.7	99.1	**91.3**	99.6	91.7
address	99.1	81.8	99.3	**87.2**	99.7	90.2
summary	99.2	50.9	99.7	**80.8**	97.5	93.8
Page number	98.1	72.9	99.9	**98.6**	99.9	92.4
press	99.4	**79.2**	99.4	76.1	99.2	88.5
college	98.8	74.9	99.4	**86.7**	99.9	92.4
title	92.2	87.2	98.9	**98.3**	99.5	70.1
volume	98.6	75.8	99.9	**97.8**	99.9	**89.2**
Average F1		77.6		**91.5**		89.7
Comprehensive accuracy	85.1%		**95.37%**		92.9%	

5. Conclusion

The conditional random field model has great competitiveness in POS tagging, natural language processing and other fields, but its performance largely depends on the parameter estimation method. A good parameter estimation algorithm is very important for the conditional random field model. This paper uses the improved particle swarm optimization algorithm to estimate the maximum likelihood parameter of the conditional random field model. And

compared with HMM and the traditional conditional random model, the conditional random field model trained by the improved particle swarm optimization algorithm has the better word accuracy and F measurement value.

In order to prevent the particle swarm optimization algorithm being fallen into local convergence in early period, the paper introduces the particle swarm's aggregation to control the convergence of the algorithm. To make the particle swarm restrain from being trapped into an infinite iteration in the optimal position, we introduce an iterative logarithmic likelihood ratio as the stop criterion. And to get the best search performance of the particle swarm optimization, we adopt the adaptive change method to acquire the inertia factor and learning factor.

References

[1] J Lafferty, A McCallum, F Pereira. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* Proc. 18th International Conf On Machine Learning. 282-289.
[2] AR Kinjo, F Rossello, G Valiente. Profile Conditional Random Fields for Modeling Protein Families with Structural Information. *BIOPHYSICS*. 2009; 5: 37-44.
[3] B Settles. *Biomedical named entity recognition using conditional random fields and novel feature sets.* In Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004). 104–107.
[4] M Bundschus, M Dejori, M Stetter, V Tresp, HP Kriegel, Extraction of semantic biomedical relations from text using conditional random fields, BMC Bioinformatics 9 (2008): 207.
[5] Fletcher R and CM Reeves. Function Minimization by Conjugate Gradients. Comp. 1964; 7:149-154.
[6] M Al-Baali. Descent property and global convergence of the Fletcher-Reeves method with inexact line search. *IMA Journal of Numerical Analysis*. 1985; 5: 121-124.
[7] YH Dai and Y Yuan. Convergence properties of the Fletcher-Reeves method. *IMA Journal of Numerical Analysis*. 1996; 16: 155-164.
[8] Charles A Sutton. Efficient Training Methods For Conditional Random Fields. PhD thesis, University of Massachusetts Amherst. 2008.
[9] Richard H Byrd, Peihuang Lu, Jorge Nocedal and Ciyou Zhu, A Limited Memory Algorithm For Bound Constrained Optimization. *SIAM Journal on Scientific Computing*. 1995; 16: 1190-1208.
[10] Thomas G Dietterich, Guohua Hao, Adam Ashenfelter. Gradient Tree Boosting for Training Conditional Random Fields. *Journal of Machine Learning Research*. 2008; 9: 2113-2139.
[11] C Sutton, A McCallum. *Piecewise pseudolikelihood for efficient training of conditional random fields*. Proceedings of the 24th international conference on Machine learning. 2007; 863-870.
[12] SVN Vishwanathan, NN Schraudolph, MW Schmidt, KP Murphy. *Accelerated training of conditional random fields with stochastic gradient methods*. Proceedings of the 23 rd International Conference on Machine Learning, Pittsburgh, PA. 2006; 969-976.
[13] Chris Pal, Charles Sutton, and Andrew McCallum. Sparse forward backward using minimum divergence beams for fast training of conditional random Fields. *In International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 2006; v-v.
[14] J Kennedy and RC Eberhart. Particle Swarm Optimization. Proc. on feedback mechanism *IEEE Int'l. Conf. on Neural Networks, IEEE Service Center*. 1995; VI: 1942-1948.
[15] Y Guangyou. *A modifed particle swarm optimizer algorithm*. 8th International Conference on Electronic Measurement and Instruments. 2007; ICEMI '07: 2-675-2-679.
[16] H anH, G iles C, M anavog lu E, et al. *Automatic Document Metadata Extraction Using Support Vector Machines*. In: Proceeding s of Joint Conference on Digital Libraries. 2003: 37 - 48.
[17] Lafferty J, M cCallum A, Pereira F. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. In: Proceedings of the 18th International Conference on Machine Learning. 2001: 282 - 289.
[18] Byrd RH, Nocedal J, Schnabel RB. Representations of Quasi-Newton Matrices and Their Use in Limited Memory Methods. *Mathematical Programming*. 1994; (2): 129- 156.
[19] Darroch JN, Ratcl iff D. Generalized Iterative Scaling for Log- linear Models. *Annals of Mathematical Statistics*. 1972; 43(5): 1470 - 1480.
[20] Della Pietra S, Della Pietra V, Lafferty J. Inducing Features of Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1997, 19(4): 380- 393.
[21] Peng F, M cC allum A. Accurate Information Extraction from Research Papers Using Conditional Random Fields. *Information Processing & Management*. 2006; 42(4): 963- 979.
[22] Sha F, Pereira F. *Shallow Parsing with Conditional Random Fields*. In: Proceeding s of Hum an Language Technology NAACL. 2003: 134-141.
[23] Ruggieri S. Efficient C4.5. *IEEE Transactions on Knowledge and Data Engineering*. 2002, 14(2): 438-444.
[24] Zave P. Classification of Research Effort s in Requirement's Engineering. *ACM Computing Surveys*. 1997; 29(4): 315-321.

[25] Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*. 2004; 5(10): 1205−1224.
[26] S Das. *Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection*. In: Proceedings of 18th Conference on Machine Learning. 2001: 74-81.
[27] Purwoharjono, Abdillah, Muhammad, Penangsang, Ontoseno, Soeprijanto, Adi. Optimal placement and sizing of thyristor-controlledseries-capacitor using gravitational search algorithm. *Telkomnika*. 2012, 10(5): 891-904.
[28] Sari, Lydia. Effects of puncturing patterns on punctured convolutional codes. *Telkomnika*. 2012; 10(4): 752-762.