ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(28), DOI: 10.17485/ijst/2015/v8i28/73900, October 2015* ISSN (Online) : 0974-5645

An Improved Alias Classification using Logistic Regression with Particle Swarm Optimization

M. Subathra[1*] and R. Nedunchezhian[2]

[1]Department of Computer Applications, PSG College of Technology, Bharathiar University, Coimbatore – 641004, Tamil Nadu, India; subaclient@gmail.com
[2]KIT-Kalaignar Karunanidhi Institute of Technology, Coimbatore – 641402, Tamil Nadu, India; rajuchezhian@gmail.com

Abstract

An improvement in detection of alias names of an entity is an important factor in many cases like terrorist and criminal network. In this paper, the social network properties are used to construct a feature set for classification. The proposed particle swarm optimization is used to optimize the regularization parameter of the logistic regression and improve the accuracy of the entity alias classification significantly to 4.98% compared to that of the logistic regression. The experimental results demonstrated its performance and the results are compared to the logistic regression with alias detection dataset.

Keywords: Alias Classification, Logistic Regression, Particle Swarm Optimization, Regularization

1. Introduction

Name disambiguation is the process of resolving the conflicts that arise when the identity of an entity is ambiguous. The lexical and referential ambiguities are the basic type of name ambiguities. The lexical ambiguity arises, when several persons have the same name. The referential ambiguity arises, when the person is referred to by many names (aliases). Bhat et al.[1] enriched Latent Semantic Analysis and Holzer et al.[2] proposed geodesic based shortest path algorithm, Paul et al.[3] investigated many classification algorithms such as K-Nearest Neighbor, decision tree, SVM and logistic regression and found that logistic regression have better performance compared to other techniques for alias detection. The pattern-based approach[4–6] with n-gram technique is used on alias extraction of an entity. Pantel et al.[7] proposed cosine similarity measure for automatically detecting aliases in malicious environments. Shen et al.[8] investigated the link- based properties for alias detection using fuzzy set based absolute order-of-magnitude model. Yin et al.[10] used the e-mail dataset specifically for extracting the e-mail user aliases. Ning et al.[11] proposed probabilistic based logistic regression to find semantic alias detection of an entity.

2. The Proposed Technique

Logistic Regression (LR) is a discriminative model of Machine Learning[12] that uses an independent variable (x) to predict the dependent variable (y). The LR classifier uses the logistic function to find the learning model. The PSO algorithm[13] is used to solve the unconstrained optimization problem with the objective function. Maximize LR classification accuracy f(x) = XεRn, where n is a dimension of X. In this application, 'n' is the number of similarity measures and 'X' is the given dataset. The classification performance of the logistic regression is maximized by the optimization of the regularization parameter λ(lambda) through the equation $\sum_{j=1}^{m} \theta_j^2$, where

* *Author for correspondence*

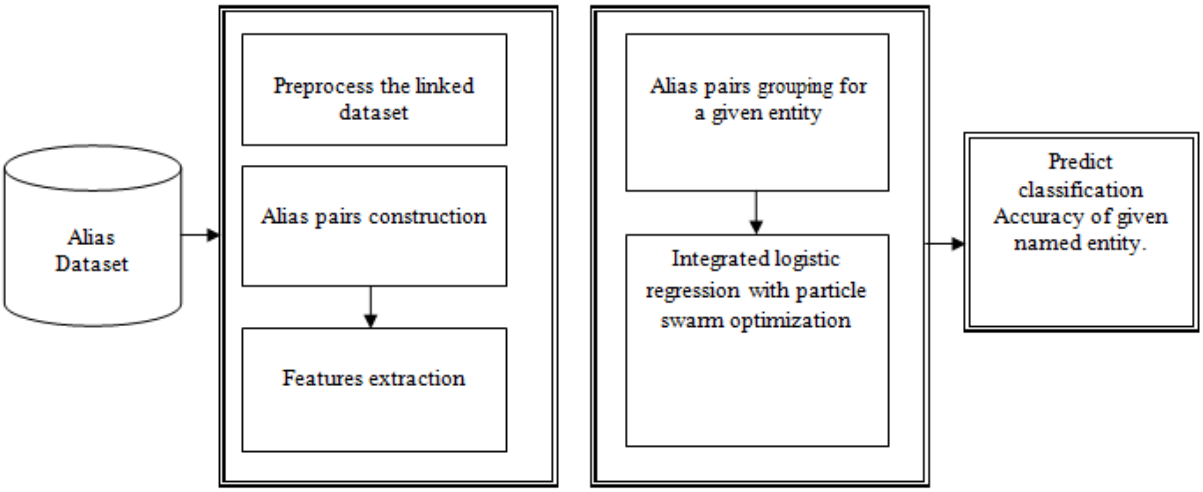


Figure 1. The Proposed Alias Detection System.

'm' is the total number of negative and positive instances and j is the weight of each feature in the learning model. Also, each name pair is associated with five features such as common friends, dot product, normalized dot product, co-occurrence relevance and social relevance which are selected and verified based on the related works[3,11]. These features are integrated and transferred to the classifier to build a learning model. The PSO is integrated to optimize the regularization parameter with the logistic regression. In regularization, all the features are taken, and its magnitude reduced to find the hypothesis function h (θ). Each feature consists of f (θ) theta to estimate the hypothesis function. To regularize these features, the regularization parameter (λ) is introduced avoiding the over fitting problem. The proposed alias detection system consists of two steps: Firstly, the alias preprocessing step transforms aliases and the real name into name pairs. Secondly, the link based features[3,11] are applied to the name pairs, and the semantic similarity values are calculated. Finally, the alias pairs are grouped according to the given entity and then, the integrated PSO with logistic regression is applied for alias classification of a given entity. The proposed work of alias detection is shown in Figure 1.

λ is regularization parameter in the classifier used to handle unbalanced dataset to avoid the over-fitting problem[14–16]. The proper value of the regularization parameter improves the learning rate of the model. Typically, threshold t = 0.5 is applied to determine whether an examined entity name pair belongs to class 1 or 2. In the training phase, the algorithm tries to solve an optimization problem with the use of PSO. It is fast in convergence and used to find the regularization parameter of logistic regression through its fitness function that has generated different lambda values by random population. The λ (particle) values of the logistic regression are randomly generated using PSO[17] to achieve an optimal value which maximizes the classification accuracy as shown in Figure 2.

Input: Data set

$DN = \{(X\ ,Y), Dimesion\ of\ X\ is\ m = 1, \ldots\ldots,5;\}, w = 1,\ L_c = .05, L_s = .0$

where $1 \le i \le P$, $1 \le j \le m$, r1, r2(random numbers).

Compute: Global best (gbest) solution

Output: The regularization parameter (plambda).

Step1: Initialize the swarm(P) of particle(plamda)s with random positions and velocity.

Step2: Repeat

for each particle plambda, in P do

find out fitness of particle in m variables

fp = f(plambda,)

If fp is better than f(pbest)

pbest = plambda,

end

gbest = best pbest, in P(swarm size)

Step3: for each particle plambda in P do

updatevelocity

$v_{ij}(t+1) = v_{ij}(t) * w + L_c * r1 * \left(plamda_{im}(t) - plamda_{ij}(t)\right) L_s * r2 * (plamd$

update position $plamda_{ij}(t+1) = plamda_{ij}(t) + v_{ij}(t+1)$

end

Figure 2. PSO for LR Regularization.

Table 1. Performance evaluation of sample entities

		Classification Accuracy		F-score		Precision		Recall	
S No	Entity Name	LR	PSO-LR	LR	PSO-LR	LR	PSO-LR	LR	PSO-LR
1	Abu-Baker	87.73	91.66	64.65	68.92	65.28	74.28	66.65	64.28
2	abu_fatima	90.65	95.63	64.51	74.26	63.51	66.12	65.69	84.62
3	abdallah_fazul	84.02	89.34	65.33	68.52	60.72	74.61	65.23	65.30
4	Hajj	93.90	97.25	65.03	74.48	62.04	65.39	68.32	86.51
5	abd_al_wakil_al_masri	87.57	91.71	64.61	68.72	65.40	74.47	65.87	63.80
6	abdul_rahman_s._taha	88.75	97.04	55.44	69.36	63.62	73.83	63.81	65.40

The size of the swarm is P (number of particles) and the dimension of the particle 'm'. The i^{th} particle is denoted asp (i) = {plambdai1, plambdai2,…, plambdaim}. The best personal position, best global position and velocity (V) at the iteration (t) are represented as pbest (t) {p_{i1}, p_{i2} ..., p_{im}}, gbest(t) = (p_{g1}, p_{g2}, p_{g3}…, p_{gn}) and Vi = {v_{i1}, v_{i2},…, v_{im}}. W is an inertia weight, L_c Cognitive and L_3 Social learning factors and r1 and r2 are the random numbers, v_{ij} (t) is the velocity of the i^{th} particle in the t^{th} iteration of j^{th} dimension, x_{ij}(t) is the j^{th} position of the i^{th} particle in the t^{th} iteration.

The dimension of the particle is an integrated feature set. The maximum iteration is 10000 and the swarm size is fixed as 20 and the number of lambda value is generated to increase the classification accuracy.

The fitness function of each particle is computed using randomly selected instances and finding the standard deviation of the population. The global lambda value is calculated, using the fitness function and the following steps are used to find the optimized parameter (λ) regularization value applied to logistic regression.

3. Results and Discussion

This section gives details of the experimental evaluation of the performance of the proposed system.

3.1 Dataset

The experiment for the proposed work is conducted on a core i5 with 4 GB RAM running on windows 7 OS. The dataset for the experiment is generated from Auton lab, CMU's School of Computer Science. It consists of three files, namely Alias, Names, and Link. The Alias Dataset[3] consists of the names of the terrorists and their aliases; the Name file, all names and the Link file, 5581 links. The alias pairs are created by incorporating both the positive and negative pairs that are around 15000. Then, the association between the names is calculated, using the link based features, and the dataset is formed and normalized before applying to the classifier. Each link contains two or more names and represents an observed relation between the names which appear together.

3.2 Performance Measure

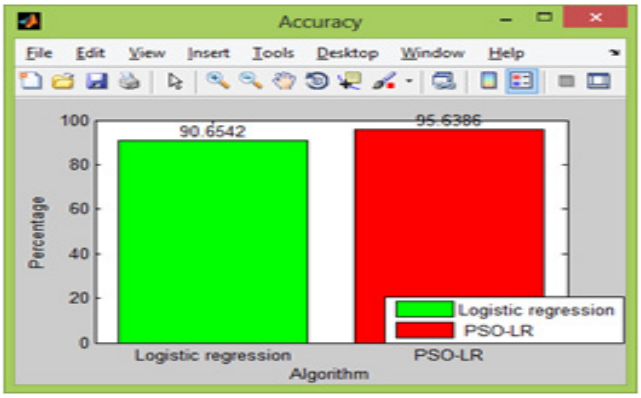
The F-Measure is used to evaluate the performance of the proposed approach that is the harmonic mean between the precision and recall.

$$\text{Precision} = \frac{\text{Number of correct aliases retrived by an algorithm}}{\text{Number of aliases retrived by an algorithm}}$$

$$\text{Recall} = \frac{\text{Number of correct aliases retrived by an algorithm}}{\text{Number of total aliases of an entity}}$$

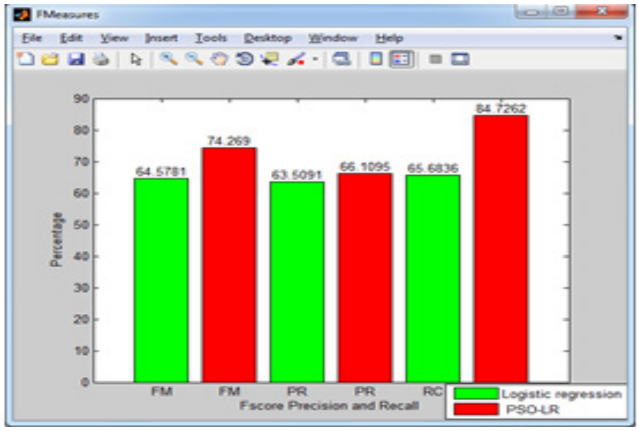
$$\text{Fscore} = 2 * \left(\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \right)$$

The experimental results are evaluated for sample entities. Figures 3(a) and 3(b) illustrate the performance analysis of the given entity of Abu-Fatima implemented in MAT lab version7.0.



(a)

Figure 3. (a) Comparison Classification Accuracy of Sample Entity.



(b)

Figure 3. (b) Comparison of F-Measure Performance of Sample Entity.

Figure 3 (a) provides the comparison of the classification accuracy between the logistic regression and the PSO-LR. The proposed method PSO-LR provides better results. The logistic regression provides an accuracy of 90.65%, and the proposed method PSO-LR gives the improved accuracy of 95.63%. Figure 3 (b) illustrates the comparison between the f-score, precision and recall. The proposed method of PSO–LR provides a better result of 74.26%, 66.12% and 84.72% for f-score, precision and recall, respectively, compared to the logistic regression result of 64.51%, 63.51% and 65.69%.

The above analysis is carried out using different entities in the dataset and the classification accuracy of the proposed system is verified and shown in Table 1. The alias detection of sample entity 'abdallah-fazul' is shown in Table 2.

Table 2. Detected alias names of a sample entity

Real Name	'abdallah_fazul'
Alias Names	'abdallah_mohammed_fazul'
	'fazul_abdalla'
	'fazul_abdilahi_mohammed'
	'fazul_abdullah_mohammed'
	'fazul_adballah'
	'fazul_mohammed'
	'haroon'
	'harun'

The learning time performance in terms of the cost function of logistic regression for positive and negative pairs is taken for 150 and 130 iterations respectively. In PSO approach, it is fast in convergence and the fitness function reduced the learning curve by 27 and 32 iterations for positive and negative pairs respectively. The results showed that the execution time of PSO based approach has performed better compared to the logistic regression for a given sample entity as shown in Figure 4.

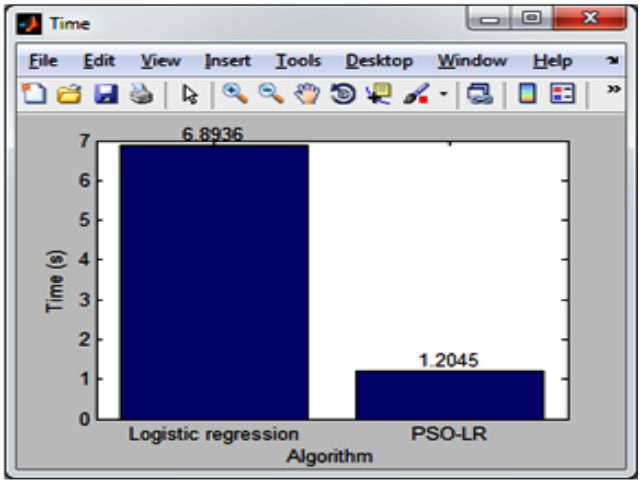


Figure 4. Execution Time Comparison of Sample Entity.

4. Conclusion

In this work, the proposed PSO is used to optimize the regularization parameter of the logistic regression and improve the alias classification accuracy of the entity significantly 4.98% compared to that of the logistic regression. The experimental results demonstrated its performance and the results are compared to the logistic regression. In future, this work may be extended to apply PSO for other classifiers also.

5. References

1. Bhat V, Oates T, Shanbhag V, Nicholas C. Finding aliases on the web using latent semantic analysis. IEEE Data and Knowledge Engineering. 2004; 49(2):129–43.
2. Holzer R, Malin B, Sweeney L. Email alias detection using social network analysis. Proceedings of the 3rd International Workshop on Link Discovery: 2005. p. 52–7.
3. Paul H, Moore A, Neill D, Schneider J. Alias Detection in Link Data Set. Proceedings of the International Conference on Intelligence Analysis; Boston; USA. 2005. p. 1–8.
4. Hokama T, Kitagawa H. Extracting mnemonic names of people from the web. Proceedings of ninth International Conference. Asian Digital Libraries; Berlin, Germany: Berlin Heidelberg. 2006. p. 121–30.

5. Anwar T, Abulaish M, Alghathbar K. Web content mining for alias identification: A first step towards suspect tracking. International Conference on Intelligence and security Informatics (ISI); Beijing, China: IEEE. 2011. p. 195–97.
6. Bollegala D, Matsuo Y, Ishizuka M. Automatic Discovery of Personal Name Aliases from the Web. IEEE Transaction on knowledge and data engineering. 2011; 23(6):831–44.
7. Pantel P. Alias Detection in Malicious Environments. Proceeding of AAAI Fall Symposium. Capturing and using Patterns for Evidence Detection; 2006. p. 1–7.
8. Shen Q, Boongoen T. Fuzzy Orders-of-Magnitude-Based Link Analysis for Qualitative Alias Detection. IEEE Transaction on knowledge and data engineering. 2012; 24(4):649–63.
9. Yin M, Luo J, Cao D, Liu X, Tan Y. User Name Alias Extraction in Emails. I.J. Image, Graphics and Signal Processing. 2011; 3(9):1–9.
10. Yin M, Xiaonan L, Luo J, Luo X. A System for Extracting and Ranking Name Aliases in Email. Journal of Software. 2013; 8(3):737–45.
11. Ning A, Jiang L, Wang J, Luo P, Wang M, Li BN. Towards detection of aliases without string similarity. Information Sciences. 2014; 261(10):89–100.
12. Andrew YNg, Feature selection, L1 vs. L2 regularization, and rotational invariance. Proceedings of the Twenty-first International Conference on Machine Learning (ICML'04); New York, USA. 2004. p. 78.
13. Kennedy J, Eberhart RC. Particle swarm optimization. Proceedings of IEEE International Conference on Neural Networks; Piscataway, New Jersey. 1995. p. 1942–48.
14. Lei X, Hu Q, Kong X, Xiong T. A Regularization Blind Image Restoration Technique by using Particle Swarm Optimization. Proceedings of the 3rd International Conference on Multimedia Technology; Guangzhou, China. 2013. p. 984–92.
15. Hlosta M, Striz R, Kupcik J, Zendulka J, Hruska T. Constrained Classification of Large Unbalanced Data by Logistic Regression and Genetic Algorithm. International Journal of Matching Learning and Computing. 2013; 3(2):214–18.
16. Lee Su-In, Lee H, Abbeel P, Ng AY. Efficient L1 Regularized Logistic Regression. Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06); London; UK. 2006; 21(1):401–08.
17. Yazd HGH, Arabshahi SJ, Tavousi M, Alvani A. Optimal Designing of Concrete Gravity Dam using Particle Swarm Optimization Algorithm (PSO). Indian Journal of Science and Technology. 2015; 8(12):1–10.