International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) ISSN: 2395-5325
Volume 3, Special Issue 3 (September '2017)
Proceeding of - International Conference on Recent Trends in Computational Life Science and Information Technology,
Conference held at Tiruppur Kumaran College for Women,Tirupur,Tamilnadu, India.

A STUDY ON EMAIL MANAGEMENT USING FEATURE SELECTION METHODS IN CLASSIFICATION ALGORITHMS

S.Divya,
Research Scholar,
Department of Computer Science,
SNR Sons College,
Coimbatore, Tamilnadu, India.

G.Maria Priscilla,
Professor & Head,
Department of Computer Science,
SNR Sons College,
Coimbatore, Tamilnadu, India.

Abstract: Email Classification is one of the vital problems in the email management due to its impact on the usage. Despite of several applications like messengers such as watsup, kaizala and social media networks such as Facebook and twitter, importance of email was kept exploring . In order to increase the performance of management it has become mandatory to automate the classification of the email against relevant and irrelavent emails This paper investigates the classification algorithms used to classify the email in terms of feature selection methods like genetic algorithm, simulated annealing and principle component analysis. This study discussesabout the importance of the classification modelsagainst the accuracy and security measures. To handle the implication on this study, we propose a novel deep learning based classification algorithm named as EmailGrading. In this feature are extracted from the low level feature in order tomaintain the hierarchical representation. Also it disentangles the abstraction on the different layers in order to improve the performance in terms of labelling and accuracy.

Keywords: Email Classification, Feature Selection, Deep Learning, Data Abstraction

I.INTRODUCTION

Email is one of the fastest and popular medium of communication [1]. It has been used extensively by millions of peoples in their everyday life as it healthy conservation around the world and it is been used as document delivery system and document archive [2]. Unlike it usage, it lead to many issue in the email management in terms of handling spam mail and denial of service attacks [3]. Though many manual solution is been proposed to handle the attacks but it has become mandatory to automate the classification of the email [4]. This paper investigates the classification algorithms used to classify the email in terms of feature selection methods like genetic algorithm, simulated annealing and principle component analysis. This study discusses about the importance of the classification models against the accuracy and security measures [5]. To handle the implication on this study, we propose a novel deep learning based classification algorithm named as EmailGrading. In this feature are extracted from the low level feature in order to maintain the hierarchical representation. Also it disentangles the abstraction on the different layers in order to improve the performance in terms of labelling and accuracy. The rest of the section is organized as follows, section 2 describes the review of literature followed by section 3 to define the proposed methodology as outline and finally section 4 concludes the study of the paper.

II.REVIEW OF LITERATURE

The review of the literature is analysed in terms of feature selection methods and its importance in the email management.

1.1. Feature selection methods

Feature Selection methods is employed to reduce size of the training model , and select the features of the importance as modelling feature set or feature subset. The following feature selection methods yields best results,

2.1.1. Genetic Based Feature Selection method for spam mail detection

In this literature Feature selected using genetic algorithm is considered as chromosomes [6]. Initially feature are generated randomly, feature will be selected based on the binary condition, i.e if binary value of the text =1, the feature is selected else feature is removed. The value is considered as fitness value. The important operation feature selection based on genetic is followed by crossover and mutation. The Crossover is used to recombine the feature selected. Mutation is to determine the capability of the feature against the transformation in the next iteration. In this feature is considered as Spam.

2.1.2. Feature Selection method based Simulated annealing

In this literature, feature selection method is analysed for identifying features which are important for each class. This entails selecting the features specifically for each class. This is carried out by using the simulated annealing technique. The algorithm is run separately for each class resulting in the feature subset for that class[7]. The simulated annealing procedure is to the iterate K times if k is the number of classes. In each iteration, the feature subset for the ith class is found. When carrying out the evaluation of the current string in the SA, all the patterns of the ith class are taken as belonging to one class and all the other patterns belong to the other class. If there are d features, the current solution of the simulated annealing has d elements where each element is either 1 or 0. If the element j is 1, it means that the feature j is present in the feature subset and if the element j is 0, it means that the feature j is not present in the feature subset. The current solution is evaluated by finding the classification accuracy of the verification set using the feature subset selected in the current solution. The SA is

used to find the best feature subset using the classification accuracy as the evaluation criterion.

2.1.3. Feature Selection Method based Principle Component Analysis

PCA is a standard statistical technique that can be used to reduce the dimensionality. PCA is a method of transforming the initial data set represented by vector samples into a new set of vector samples with derived dimensions [8]. The transformation is based on the assumption that high information corresponds to high variance. The first principal component minimizes the distance of the sum of squares between data points. PCA returns linear combinations of the original features by utilizing the correlation and covariance matrix for the email messages.

2.1.4. Email Classification Based particle Swarm Optimization

Particle Swarm optimization is considered as unsupervised clustering model to classify the emails. Feature selection (FS) as a global optimization problem which decreases dimensionality and improves the accuracy of spam email classification. PSO as a computational model fallows the social behaviour of bird flocking or fish schooling. The proposed PSO-based feature selection algorithm searches the feature space for the best feature subsets. The evolution of feature selected is determined by a fitness function. The classifier performance and the length of selected feature vector as a classifier input are considered for performance evaluation.

2.1.5. Hybrid Feature Selection Algorithm based on rough Set Theory and TFIDF techniques

A combination of two types of feature selection methods named as rough Set theory and TFIDF is used in the classification task. Feature selection is used to reduce the dimensionality of word frequency without affecting the performance of the classification task [11]. However, these two techniques sometime will return poor result of classification because of inadequate data or information if they work individually. Therefore, this proposed hybrid method hope to increase the classification task accuracy rate because these two techniques will facilitate the classifier in generating a good filtering result.

2.1.6. Email Classification Attributes through feature Selection methods

In this the degree of influentially is measured by applying four data mining algorithms on a large set of features such as information gain , gini index, Fuzzy Adaptive particle Swarm optimization and Chi Square test. These techniques are used to extract the possible no of the feature for classification. Among these best performing approaches is determined using precision and recall values [12]

2.2. Importance of the classification models

A large set of personal emails is used for the purpose of folder and subject classifications. Algorithms are developed to perform clustering and classification for this large text collection. The Classification algorithm is employed due to several reasons such as spam email classification, phishing email classification, outlier detection and multi folder categorization [9]. Machine learning algorithms are used for classification of objects of different classes. Such algorithms have proved to be efficient in classifying emails as spam, phishing and as folder for email organization.

III.OUTLINE OF THE PROPOSED METHODOLOGY

Deep learning based classification algorithm named as EmailGrading is used to hierarchically extract the deep features. Deep Features can be extracted using hybrid of principle component analysis and logical regression. Deep feature can also be extracted using the markov random field. The MRF and PCA have possibility to obtain the useful high level features. This model is considered as an automatic email classifier that automatically classifies emails into one or more of a discrete set of predefined categories.

IV.CONCLUSION

In this paper, a systematic review on the email classification models was presented. The analysis is carried on basis of the feature selection methods. The genetic algorithm simulated annealing and principle component analysis has gained as important methods for feature selection as it is used to improve the classification accuracy. The feature selection method considered to be iterative process as to include the new features or terms without rebuilding the entire learning model.

V.REFERENCES

[1]. E. Blanzieri and A. Bryl, "A survey of learning-based techniques of email spam filtering,"Artificial Intelligence Review, vol. 29, pp. 63-92, 2008.

[2]. Y. W. Wang, Y. N. Liu, L. Z. Feng, and X. D. Zhu, "Novel feature selection method based on harmony search for email classification,"Knowledge-Based Systems, vol. 73, pp. 311-323, Jan 2015.

[3]. M. H. Song, "E-Mail Classification based Learning Algorithm Using Support vector machine," in Materials, Mechanical Engineering and Manufacture, Pts 1-3. vol. 268-270, H. Liu, Y. Yang, S. Shen, Z. Zhong, L. Zheng, and P. Feng, Eds., edStafa-Zurich: Trans Tech Publications Ltd, 2013, pp. 1844-1848.

[4]. J. R. Mendez, M. Reboiro-Jato, F. Diaz, E. Diaz, and F. Fdez-Riverola, "Grindstone4Spam: An optimization toolkit for boosting e-mail classification,"Journal of Systems and Software, vol. 85, pp. 2909-2920, Dec 2012.

[5]. T. S. Moh and N. Lee, "Reducing Classification Times for Email Spam Using Incremental Multiple Instance Classifiers," in Information Intelligence, Systems, Technology and Management. vol. 141, S. Dua, S. Sahni, and D. P. Goyal, Eds., ed Berlin: Springer-Verlag Berlin, 2011, pp. 189-197

[6]. SorayyaMirzapourKalaibar ,SeyedNaserRazavi Spam filtering by using Genetic based Feature Selection" in International Journal of Computer Applications Technology and Research, vol:3 issue 12, 2014

[7]. V. Susheela Dev "Class Specific Feature Selection Using Simulated Annealing" in Mining Intelligence and knowledge exploration, 3rd international conference in Springer 2015 , pp. 12–21

[8]. Juan Carlos Gomez "PCA Document Reconstruction for Email Classification" In Computational Statistics & Data Analysis, 2012

[9]. AHMAD A. AL SALLAB, MOHSEN A. RASHWAN " E-MAIL CLASSIFICATION USING DEEP NETWORKS" in Journal of Theoretical and Applied Information Technology, vol:37, issue :2

[10]. Yimimg Yang et.al "Personalized Email Prioritization Based on Content and Social Network Analysis" in IEEE Intelligent Systems in Volume: 25, Issue: 4, July-Aug. 2010

[11]. Masurah Mohamad ; Ali Selamat "An evaluation on the efficiency of hybrid feature selection in spam email classification in International Conference on Computer, Communications, and Control Technology (I4CT), 2015

[12]. IssaQabajeh ;FadiThabtah" An Experimental Study for Assessing Email Classification Attributes Using Feature Selection Methods" 3rd International Conference on Advanced Computer Science Applications and Technologies (ACSAT), 2014

