Abstract

With the rapid development of Internet, e-mail has become effective means of communication to share information. Through e-mail, we can send text messages, images, audio and video clips across the world within a fraction of time. In recent years, e-mail users are facing problem due to spam e-mails. Spam e-mails are unsolicited commercial/bulk e-mails sent by spammers. There are many serious problems associated with spam e-mails, e.g. it may contain hyperlink which may lead to a bogus website which might ask you for your personal information like username, password, bank account number etc. Spam e-mail is not only wastage of storage space but also wastage of time. In order to tackle problems faced by users due to spam e-mail, it is necessary to classify them with the help of intelligent and robust classifier. These classifiers should have the capability to classify spam e-mail against non-spam e-mail. The spam e-mail classifier performance can be greatly enhanced with the use of artificial neural network classification algorithm. An Artificial Neural Network (ANN) is a powerful tool used for classification of data, it has capability of learning huge amount of data with high dimensionality in better way, there are various parameters of ANN to be set to tune for the better performance of neural network model, these are learning rate, architecture of ANN and momentum, these all
parameters play a very important role in improving the accuracy of ANN model. In this paper Error Back Propagation Network (EBPN) techniques based on ANN are explored with different value of learning rate from 0.2 to 0.9. An EBPN model is derived from e-mail data set obtained from UCI repository site with three different partitions. Due to high dimensionality of data set, we have applied feature selection technique for the best model. This model is tested with various combinations of feature and it is concluded that model is producing highest accuracy of 98.49% on testing samples with 52 features. The derived model is also measured with precision, recall and F-measure and achieved 98.34%, 99.07% and 98.70% respectively.

References

- Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques"; Morgan Kaufmann, San Francisco, Second Edition, 2006.
- K. , J. , Cios et al., "Data mining methods for knowledge discovery"; 3rd printing, kluwer academic publishers, (USA),2000.

Index Terms

Computer Science  
Artificial Intelligence
Keywords
Spam e-mail    Classification    Error Back Propagation Network (EBPN)    Feature Selection