International Journal of Engineering & Technology, 7 (1.1) (2018) 283-285

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Analysis of supervised classification techniques

P. Lakhmi Prasanna, D. Rajeswara Rao, Y. Meghana *, K. Maithri, T. Dhinesh

Koneru Lakshmaiah Educational Foundation
**Corresponding author E-mail: meghanayenikapati@gmail.com*

Abstract

As the number of digital documents and data are being increased rapidly, it is important to classify them in to respective categories. This process of classifying the data is called classification. There are three ways in to which the data can be classified un supervised, supervised and semi supervised methods. Automatic Text Classification is done by supervised learning techniques. This paper discusses about various classification techniques, their advantages and limitations. Finally, it concludes with the best classification technique. In this paper the best classification technique that was proposed is Artificial Neural Network (ANN). The reason for proposing ANN as the best algorithm is given and its application in various important fields was given.

Keywords: *KNN; Naïve Bayes; Support Vector; Decision Tree; ANN.*

1. Introduction

As everyone know that the amount of information available in web is huge and complex. It is a tedious task to classify millions of documents manually. So Automatic Text Classification came into existence. This classifier can be constructed by considering some pre-classified documents which are more accurate and efficient than manual text classification. Text Mining is a process of retrieving quality data which is useful and relevant to the user. Text classification is a process of classifying the documents into predefined categories.
"Natural Language Processing", "Data Mining" and "Machine Learning Techniques" work together for Automatic classification of documents. The goal of text mining is to enable users to retrieve useful data from text resources. Many techniques were proposed and being proposed by many people. But no one is able to propose a best technique among the proposed ones. It is very important that a best technique for classification should be used. In this paper, various techniques are discussed along with their advantages and limitations. And this paper finally proposes ANN as the best technique among the discussed classification algorithms by giving valid reasons.

2. Classification techniques

There are various supervised techniques for classification. Some of them are discussed below.

2.1. K-nearest neighbour (KNN)

KNN is used to classify documents based on similarity between input documents and training data. The data that was classified is stored so that the category to which the test documents belong to will be determined. This method will classify documents based on the least distance between documents in the training data set. The training data sets are depicted on multi-dimensional word space. The word space is divided into various areas depending on the category to which the training data belongs to. A point in word space will be assigned to a category if the category is the most frequent compared to the k nearest training data. In order to assign a document to a category, Euclidian distance is generally used which computes the distance between vectors. At first, feature vectors and categories of the training set are retained. Next distances among the input vector to all retained vectors are calculated and k nearest samples are selected. The category to which the document belongs is decided based on the nearest vector which has been assigned to a specific category.
Advantages: This method is simple to implement as it uses only two parameters and is robust to noisy data.
Limitations: The time needed to compute similarity or dissimilarity is more. Practically, it is impossible to implement KNN algorithm for high dimensions and huge samples. As a result, Classification cost becomes very high for KNN. Also, the classifier grows with the number of training documents.

2.2. Naïve bayes algorithm

This algorithm applies Bayes' theorem assuming strong independence between two words. Naïve Bayes classifier is the probabilistic classifier. By assuming independence among features, the order of features will be irrelevant and thereby the presence of one feature does not affect the presence of other features during classification. Since this is a probability model, these classifiers can be trained productively using comparatively less amount of training data for the estimation of parameters necessary for classification. An assumption of independence among variables was made, only dissimilarities of the features in each class should be intended instead of determining entire covariance matrix.
Advantages: These classifiers work good in several practical situations than one generally expects. It requires less amount of training data for the estimation of parameters necessary for classification.
Limitations: The disadvantage of this classifier is its comparatively low classification performance among other algorithms such as SVM. The independence relation is violated by practically collect-

Copyright © 2018 P. Lakhmi Prasanna et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ed data and perform poorly when features are highly associated and neglects the frequency of occurrences of word.

2.3. Support vector machine (SVM)

SVM is described by the "Maximum Margin Classifier" which is a speculative classifier. The numerical input variables of the data form an n-dimensional space. Let us say there are two input variables, they will form a two-dimensional space. This input variable space will be divided by a hyper plane. In SVM, hyper plane is used to separate the points of the input variable space based on their category, either category 0 or category 1. By substituting the input values in line equation, one can say a new point is above or below the hyper plane. If the input point is above the line, the equation returns a value greater than zero which means that the input point belongs to the first category. If it is below the hyper plane, the equation returns a value less than zero which says that the point belongs to the second category. If the value returned is close to zero, then the point is closer to the hyper plane and the point may be difficult to classify. The distance between the hyper plane and the closest data points is termed as margin. The best hyper plane that can separate the two categories is the hyper plane which has largest margin known as Maximum-Margin hyper plane. The margin is the perpendicular distance between the hyper plane to closest points. Only these points can define the hyper plane and are relevant for construction of classifier. These closest points are known as support vectors.
Advantages: This technique can manage high-dimensional inputs. This method is outstanding because of its effective classification.
Limitations: The disadvantage of SVM is its comparatively complex nature of training and classification algorithms and also during training and classifying it takes more time and memory. Also, there is a chance of occurrence of confusions during classification tasks since the same document can belong to different categories because of the similarity that is calculated for each category.

2.4. Artificial neural networks

Neural Network is generally called as artificial neural network as it is a part of Artificial Intelligence. ANN instructs the system to execute task, rather than programming the system to do particular tasks. Artificial Intelligence System will perform such tasks.
An artificial neural network comprises many artificial neurons that are associated together in accordance with network explicitly. The goal of the neural network is for a given input it should give significant outputs. ANNs are used in many areas such as,

- Bankruptcy prediction
- Speech recognition
- Fault detection

The practical problems that are represented by multidimensional datasets are considered from medical field. The dataset is divided in to training and testing sets where the testing data set has no usage in training process. The training set is collected from 2/3rd of the dataset and the remaining has been considered as test set.
In order to train a neural network, back propagation algorithm is used. By combining appropriate training function, learning function and transfer function the dataset classification uses back propagation neural network which is the most successful tool. The combination of "TRAINLM", "LEARNGDM" and "LOGSIG" works better for comparatively smaller datasets and the combination of "TRAINSCG", "LEARNGDM" and "LOGSIG" is better for larger datasets.
ANN is used in different applications like classification of remote sensing images [13], "classification of neck movement patterns related to Whiplash-associated disorders" [14], "Classification of breast cancer data [15], for detecting misfire in gasoline engines"[16], "classification of car seat fabrics"[12].
Advantages: Neural Networks are self-adaptable. They are universal functional approximators and non-linear models.
Since there is a problem in finding the most appropriate group of training function, learning function and transfer function for the classification of datasets with growing features and classified sets, it is applied in various important fields that are described above. Generally artificial neural networks are considered as simulated brains where one can give their own programming as they want to neurons and the behaviour of neurons reflects the human brain, but neurons can also be used to solve problems that were never considered before.

3. Literature survey

The term "Data mining" was coined in 1990s, but it is the evolution of one of the fields with long history. Data mining can be found in three families: "classical statistics", "artificial intelligence", and "machine learning". Artificial Intelligence attempts to apply "human-thought-like" processing to statistical problems. Some AI concepts which were adopted by some high-end commercial products are "query optimization modules for Relational Database Management Systems (RDBMS)". Machine learning is the combination of statistics and Artificial Intelligence. It can be considered as an evolution of Artificial Intelligence, because it combines AI heuristics and advanced statistical analysis. Machine learning let computer programs learn about the data they study, such that programs make different decisions based on the qualities of the studied data, using statistics for fundamental concepts, and adding more advanced AI heuristics and algorithms to achieve its goals. Fundamentally Data Mining is the application of machine learning techniques to business applications. Data mining is best described as the union of historical and recent developments in statistics, AI, and machine learning. These techniques are then used together to study data and find hidden trends or patterns within.

4. Conclusion

By studying and analysing different classification techniques, it is concluded that Artificial Neural Networks (ANN) is the best technique among all because of the advantages that are given above. If Maximum-likelihood method is compared to Back propagation neural network, Back propagation neural network is more accurate than Maximum-likelihood method. The same was happened with the case of neck movement pattern classification. Even though BPNN convergence is slow it is guaranteed. The implementation of BPN in parallel architectures is also easy that decreases the processing time compared to other algorithms.

References

[1] "A Review of Machine Learning Algorithms for Text Document Classification" by Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, Khairullah khan.
[2] "Survey of Text classification Algorithms" by Charu C. Aggarwal
[3] "Comparison of Text Classification Algorithms" by M. Trivedi, S. Sharma, N. Soni, S. Nair
[4] "Review on Classification Based on Artificial Neural Networks" by Saravanan K and S. Sasithra
[5] "Functional Analysis of Artificial Neural Network for Dataset Classification" by Rojal ina Priyadarshini, Nillamadhab Dash, Tripti Swarnkar, Rachita Misra.
[6] "An overview on Text Classification Techniques" by Dinesh Tharwani
[7] "Classification Using ANN: A Review" by Rajni Bala, Dr.Dharmender Kumar
[8] C. C. Aggarwal, S. C. Gates, P. S. Yu. On Using Partial Supervision for Text Categorization, IEEE Transactions on Knowledge and Data Engineering
[9] C. Apte, F. Damerau, S. Weiss. Automated Learning of Decision Rules for Text Categorization
[10] Rojalina Priyadarshini; "Functional Analysis of Artificial Neural Network for Dataset Classification".
[11] Guoqiang Peter Zhang "Neural Network for Classification- A Survey 2000" IEEE Transactions on systems, man and cybermetics-part c: applications and reviews, Vol 30

[12] "Neural Network based classification of car seat fabrics" International Journal of Mathematical Models and Methods in Applied Sciences by R. Furferi, L. Governi.
[13] E. Hosseini Aria, J. Amini, M.R. Saradjian, "Back Propagation Neural Network for Classification of IRS-1D Satellite Images" Vol.1, Issue.2, 2003.
[14] Helena Grip, Fredrik Öhberg, Urban Wiklund, Ylva Sterner, J. Stefan Karlsson, and Björn Gerdle, "Classification of Neck Movement Patterns Related to Whiplash-Associated Disorders Using Neural Networks", IEEE transactions on information technology in biomedicine, Vol.7, Issue.4,2003. https://doi.org/10.1109/TITB.2003.821322.
[15] Guoqiang Peter Zhang, "Classification of Breast Cancer Data with Harmony Search and Back Propagation Based Artificial Neural Network", IEEE 22nd Signal Processing and Communications Applications Conference, 2014.
[16] Abid Ali, Olaf Magnor and Matthias Schultalbers, "Misfire Detection Using a Neural Network Based Pattern Recognition", "International Conference on Artificial Intelligence and Computational Intelligence", Vol.2, Issue.3, 2009.