

Text Categorization Comparison between Simple BPNN and Combinatorial Method of LSI and BPNN

Hemlata Tekwani

Truba Institute of Engineering & Information Technology
Dept. of Computer Science & Engineering
RGPV University, Bhopal

Mahak Motwani

Truba Institute of Eng & Information Technology
Dept. of Computer Science & Engineering
RGPV University, Bhopal

ABSTRACT

This paper proposed a text categorization comparison between simple BPNN and Combinatorial method of LSI and BPNN. In the traditional error back propagation network, the weight adjustment process gets block in a local minima and also the training speed of such network is very slow which leads to reduced performance and reduced efficiency of the network. Also the Learning time of overall network is very high. Hence, to improve the categorization accuracy, a new combinatorial method of LSI (latent semantic Indexing) and BPNN (back propagation neural network) is proposed. The latent semantics demonstration is an accurate data structure in low-dimensional space in which documents, terms and queries are rooted and also compared. Singular value decomposition (SVD) technique is used in Latent semantic Analysis in which large term-document matrix is decomposed into a set of k orthogonal factors by which the original textual data is changed to a smaller semantic space. New document vectors are found in reduced k -dimensional space. Also new coordinates of the queries are found. Here we implement combinatorial method of LSI and BPNN based technique for the classification of 20Newsgroup dataset which include categories of Sports, CS, and Medicine. The proposed technique implemented is compared with the existing BPNN technique. Hence, this new method greatly reduces the dimension and better classification results can be achieved.

KEYWORDS

Text categorization, Latent semantic Indexing, Singular value decomposition Neural network, Back propagation neural network.

INTRODUCTION:

Today is a time of networking or usually known as Internet. The ways in which users interact with a World Wide Web (Web) site available on Internet provides enormous data processing on the usefulness and effectiveness of Web design elements and content built in it. Due to this the large amount of text are usually uploaded into many sites and thus it need to be classified. Data mining is the process of extracting useful information from databases. Many approaches to temporal data mining have been proposed to extract useful information, such as time series analysis, temporal association rules mining, and sequential pattern discovery. Several core techniques that are used in data mining describe the type of mining and data recovery operation.

Text classification is striking and remarkable because a large volume of text documents is available in many applications. Text categorization or classification assigns or classes to concealed text documents. Classes may be represented numerically or by using single word or sentence or phrase etc. Earlier approach signifies categorization of text manually using domain professionals. Thus, there is a requirement of automated text classification Scenario. It automatically assigns categories and helps in simplicity of storage and retrieval of suitable text

documents and its various contents for various real-world applications [1].

A neural network is a kind of data processing system that consisting of a huge amount of simple and highly interconnected processing fundamentals in architecture stimulated by the structure of the intellectual cortex portion of the brain. For this reason neural networks are often capable of doing things like humans or animals do well but which conventional computers often do poorly. Neural networks have come out in the past few years as an area of extraordinary chance for research development and application to various real worlds problem. Indeed neural networks demonstrate characteristics and capabilities not provided by any other technology [1].

The recent reappearance of attention in neural networks, machine learning, and parallel computation has led to rehabilitated research in the area of statistical pattern classification. Near the beginning, pattern classification research achieved in the '60s and '70s focused on infinite training data properties of classifiers on representative convergence of density estimators and on providing bounds for error rates. Many researchers considered parametric Bayesian classifiers where the form of input distributions is unspecified to be known and parameters of distributions are predictable using techniques that necessitate simultaneous access to all training data [2].

The ability to learn is a fundamental trait of intelligence. Although a precise definition of learning is difficult to formulate,

Learning process in ANN context can be described as the dilemma of updating network architecture and connection weights so that a network can efficiently perform a specific task. The neural network typically learns about the connection weights from various existing training patterns. Weights are updated iteratively in the network which increases performance and efficiency of the network ANNs' ability to automatically learn from examples makes them attractive and exciting. ANNs become skilled at underlying rules like relationships between input and output from the group of several examples, but it does not learn from set of rules specified by human experts. This is the key advantage of neural network in comparison with traditional expert systems [3].

There are three main learning paradigms: supervised, unsupervised, and hybrid. In supervised learning, or learning with a "teacher," the network is provided with a correct answer (output) for every input pattern. Weights are determined to allow the network to produce answers as close as possible to the known correct answers. Reinforcement learning is a variant of supervised learning in which the network is provided with only a critique on the correctness of network outputs, not the correct answers themselves. In contrast, unsupervised learning, or learning without a teacher, does not require a correct answer associated with each input pattern in the training data set. It explores the underlying structure in the data, or correlations

between patterns in the data, and organizes patterns into categories from these correlations. Hybrid learning combines supervised and unsupervised learning. Parts of the weights are usually determined through supervised learning, while the others are obtained through unsupervised learning [3].

Learning theory must address three fundamental and practical issues associated with learning from samples: capacity, sample complexity, and computational complexity. Capacity concerns how many patterns can be stored, and what functions and decision boundaries a network can form. Numerous efforts to develop “intelligent” programs based on von Neumann’s centralized architecture have not resulted in general-purpose intelligent programs. Inspired by biological neural networks, ANNs are massively parallel computing systems consisting of an extremely large number of simple processors with many interconnections. ANN models attempt to use some “organizational” principles believed to be used in the human [3].

A learning algorithm is an adaptive method by which a network of computing units self-organizes to implement the desired behavior. This is done in some learning algorithms by presenting some examples of the desired input-output mapping to the network. A correction step is executed iteratively until the network learns to produce the desired response. The learning algorithm is a closed loop of presentation of examples and of corrections to the network parameter.

In general, learning can be classified into two types supervised learning and unsupervised learning. In supervised training, both the inputs and the outputs are provided. The network then processes the inputs and compares its resulting outputs against the desired outputs. Errors are then calculated, causing the system to adjust the weights which control the network. This process occurs over and over as the weights are continually tweaked. In unsupervised training, the network is provided with inputs but not with desired outputs. The system itself must then decide what features it will use to group the input data. This is often referred to as self-organization or adaption.

Semi-supervised learning can be applied when limited amount of training data is accessible. Because semi-supervised learning requires less human effort and Gives higher accuracy, it is of great advantage both in theory and in practice. In many Classification Applications labeled Training data are scarce but unlabeled data are plentiful. It is very usable if we can use unlabeled data to aid labeled data in learning a classifier. Semi Supervised learning deals with such problems. Some representative semi-supervised learning methods include Mixture model, EM, Transductive SVM, Cotraining, Graph Methods.

LSI (Latent semantic indexing) has been applied to text categorization technique. For the process of retrieval of information to decipher problems of synonymy of various words of natural language, Latent semantic indexing (LSI) is used. The latent semantics representation is an precise data structure in short-dimensional space in which documents, terms and queries are deep-rooted and are compared to understand and solve the problem of synonymy of words. Latent semantic Analysis uses singular value decomposition (SVD) technique in which large term-document matrix is changed into a set of k orthogonal factors so that the unique textual data is converted to a reduced semantic space. Also, new document vectors and new coordinates of the queries are calculated and compared in smaller k -dimensional space [4].

LITERATURE SURVEY:

In the year 1993, Jennifer Farkas proposed “Documents, Concepts and Neural Networks” [6]. In this paper the relevance of neural networks to the problem of document classification is investigated. The basic method is to identify sets of keywords in a given set of documents, to group these keywords into semantic similarity classes, which are called concepts, and to assign classifiers, according to the conceptual patterns they contain. The back-propagation learning algorithm is applied to the textual documents which are represented numerically in a semantically meaningful way to build a document classifying neural network. The network can be taught to classify natural language text according to predefined specifications.

The network NeuronZ uses the delta learning rule for the purpose of updating the weights of the connections between the processing elements, and to reduce the difference (the delta) between the desired output value and the actual output value of a processing element during a training cycle.

This paper shows that neural networks provide a new platform for the automatic classification of semantically similar documents with the convergence property of prototype NeuroZ. Hence a better system can be built which distinguishes between relatively complex linguistic patterns and resulting system can classify wider range of documents relative to more heterogeneous topics.

In the year 2007, Ram Dayal Goyal proposed “Knowledge Based Neural Network for Text Classification” [7]. In this paper, A Text Classification approach that combines Naive Bayesian text classification technique and neural networks is proposed. The author use Naive Bayesian classifier at first which has the linear separating surfaces, then the separating surfaces are modified using neural network to find better separating surfaces and hence better classification accuracy over validation data can be achieved .

The Bayesian method is highly appropriate for text classification but it suffers from a number of problems. When there is large number of categories, lack of uniformity in training data becomes a big problem. Presence of noisy data outliers creates many problems. Also, when documents are very small, the problem becomes more arduous. Some nodes may get less training documents, while other may get a very large number so these problems are solved in the proposed technique.

Sometimes, a large volume of training data is available but no prior knowledge is available and sometimes, a strong prior knowledge is available but training data is not sufficient. Thus, this paper focuses on the systems that take prior knowledge as an explicit input and learn using a combination of inductive as well as analytical learning. This paper focuses on learning problems which vary by availability of training data and prior knowledge. Basically, there are two types of learning

Combination of “Analytical” and “Inductive” learning gives more advantages so the author has applied the same technique for text classification purpose as well. A neural network is constructed with initial weights learned using Naive Bayesian classifier. This neural network starts with an accuracy of Naive Bayes. It gives improved accuracy.

In the year 2006 Zhihang Chen, Chengwen Ni and Yi L. Murphey proposed “Neural Network Approaches for Text Document Categorization” [8]. This paper proposes research in text document categorization using a categorical neural network system and a hierarchical neural network system A single neural network not sufficient to provide good classification or efficient

training because in text document categorization the feature spaces have high dimensions, training data are large and the categories are many so the author present a hierarchical neural network system and a categorical neural network system for document classification because it gives the highest accuracy in document categorization.

This paper shows an application in engineering diagnostic document categorization that the two proposed systems are more operative, effective and efficient than a single neural network, and the hierarchical neural network system gives the utmost accuracy in text document categorization.

Neural networks, Support vector machines, Genetic programming, Kohonen type self-organizing maps, Self-organizing maps, Fuzzy k-mean, Hierarchical Bayesian clustering, Bayesian network classifier and Naive Bayes classifier are machine learning techniques which have been actively explored for text document classification and categorization. Such techniques when applied in text document categorization enormous training data are found, diverse categories are found and also feature spaces have soaring dimensions. Due to this combination system training becomes highly complicated and it takes long training time and also lowers system performances.

In terms of system performances including categorization accuracy, training time and query time, It is shown that hierarchical and categorical system approaches are better than the single neural network approach .The hierarchical approach gives a significant lead in document categorization accuracy, and reasonable training and query time, therefore it is best suited by text document categorization when high dimensional feature spaces and a large number of categories are involved.

In the year 2006, Zhan Wang, Yifan He and Minghu Jiang proposed “A Comparison among Three Neural Networks for Text Classification” [9]. This paper examines the effectiveness of three neural networks, the Competitive network, the Backpropagation (BP) network and the Radial Basis Function (RBF) network in text classification. The three networks are used for Pattern classification and function approximation. The Competitive Network is a kind of unsupervised learning which is used in data clustering. The BP network and RBF network are widely used among artificial neural networks for automatic text classification.

Text classification is a process to decide the category automatically based on the content of the texts. It is necessary to categorize huge amount of data available on web server. Fundamentally, feature selector and text classifier are two parts used for text classification. Feature selector reflects to set of features and the classifier classifies the feature vector to the category it belongs. Many techniques can be used in feature selection to improve accuracy and to reduce the dimension of the feature vector and hence the time of computation will be minimized.

In the Year 2007, Cheng Hua Li and Soon Cheol Park proposed “Neural Network for Text Classification Based on Singular Value Decomposition”[10]. In this paper a new text classification models based on artificial neural networks (ANN) and Singular Value Decomposition (SVD) has been proposed.

Training of the neural network is done by Multi-Output Perceptron Learning Algorithm (MOPL) and Back-Propagation Neural Network (BPNN). Since the information provided by many labeled samples is very limited, Hence Singular value decomposition technique is introduced and used in this paper.

Singular Value decomposition is used for the extraction of features from large set of data.

The introduction of Singular value decomposition overcomes the problem existing in commonly used vector space model method for text representation. The proposed methodology reduces the vector spaces and training time in both BPNN and MOPL. The result analysis shows that the proposed algorithms has better performance as compared to the existing SVD systems.

In the Year 2003, Tim Andersen, Wei Zhang proposed “Features for Neural Net Based Region Identification of Newspaper Documents” [11]. Tim Anderson and Wei Zhang proposed a new technique of Newspaper document and non region text identification using features of neural network. In this paper, neural network based approach is used for region classification which is a significant step in document image analysis of newspaper documents.

Document image analysis is used to categorize text, graphics and pictures in images and extract useful information out of it. There are two categories of document image analysis which are Textual processing which deals with Text components and non-textual processing which deals with the non-text components of a document image. Basic Goal in textual processing is to divide the content into headline components and paragraph components and then link all the components together. For non-textual components separation is done into graphics and photos .

This paper specifies features for non-text region identification which helps in analysis and processing of image. Also, it is proved that with multiple output nodes in the neural network, Region classification can be done with better accuracy. Classification accuracy of this proposed algorithm can also be enhanced by using contextual information and features of the adjacent regions.

In the year 2011 Suvad Selman, Alma Husagic-Selman proposed “Multilayered Feed forward Neural Networks as a Tool for Distinction of the Authors of Texts” [12].In this paper, multilayered feed forward neural network is used to identify the author of a text. The algorithm is based on feed forward multi layer neural network. The multi layered network consists of a number of input layers and output layers. These layers are used for the distinction of author’s text in the document. The computational complexity is solved by dividing it into a number of computationally simple tasks where the input space is divided into a set of subspaces and then combining the solutions to those tasks.

In this paper, Author focuses on the principal aim of author distinction which is to extract uncertainty about the author of some text, which can be used in literary tasks of textual analysis for works edited, translated, with disputed authorship or anonymous, but also with forensic aspect in view to detect plagiarism, forgery of the whole document or its constituent parts, verify ransom notes, etc.

In this paper it is described that there are used style markers which are used to find writer invariants which are based on textual Properties belonging to either of four categories: lexical, syntactic, structural, and content-specific. Lexical descriptors provide statistics of total number of words or characters, average number of words per sentence, characters per sentence or characters per word, frequency of usage for individual letters or distribution of word length. Syntactic features reflect the structure of sentences, which can be simple or complex, or conditional, built with punctuation marks. Structural attributes express the organization of text into paragraphs, headings, signatures, embedded drawings or pictures, and also special font

types or its formatting that go with layout. Content-specific properties recognise some keywords: words of special meaning or significant importance for the given context.

This paper concerning author identification analysis shows efficiency of an Artificial Neural Networks when they applied in classification tasks. The proposed algorithm implemented here is used for the distinction of author's text using neural network and the technique used for the feature extraction is efficient in terms of error rate and training rate and also the time computation is less.

In the year 2010, Qiuming Ma, Zhiguang Qin, Fengli Zhang, Qiao Liu proposed "Text Spam Neural Network Classification Algorithm" [13] Here in this paper text spam classification using neural network is proposed. Email spam, also known as junk email or unsolicited bulk email (UBE), is a subset of electronic spam involving nearly identical messages sent to numerous recipients by email. Clicking on links in spam email may send users to phishing web sites or sites that are hosting malware. Spam email may also include malware as scripts or other executable file attachments. Definitions of spam usually include the aspects that email is unsolicited and sent in bulk. One subset of UBE is UCE (unsolicited commercial email). The junk email problem is rapidly becoming unmanageable, and threatens to destroy email as a useful means of communication ensure a correct mapping between the features and the classes.

This work presents a supervised learning algorithm based on Neural Network. Even though some AIS for classification have been developed, this model has a unique feature; it includes the local feature selection mechanism. The aim of feature selection is to reduce the dimension of the input vector by the selection of a feature (variable) subset which describes the object in the best manner and ensures the best Quality of the learning model.

The training corpus is splitted into a number of feature vector space so that during the testing of these spams the features can be classified. Also, cross validation methods are used to examine simulation results of the network simulation results of the network.

In the year 2011, Diganta Saha proposed "Web Text Classification Using a Neural Network" [14]. Here, an efficient implementation of web text classification system is proposed with the help of neural network technique. This classification technique is used for classification of English text documents, depending on the content of these documents. The Classifier classifies the documents into a number of classes whose number is not known. The neural network specified here consists of input layers and an output layer. The input layers are equal to the number of features in the input pattern and the output layer is equal to the number of clusters present other output nodes and the remaining input unlabelled nodes.

The proposed algorithm implemented here is used to classify given text documents into different classes and here it is specified that the total number of classes existing in the dataset is not known a priori.

The proposed algorithm implemented here provides 100% accuracy rate when the dataset taken is very large.

In the year 2009, Jingwen Tian, Meijuan Gao, and Yang Sun proposed study on classification techniques using fuzzy systems [15]. In this paper, first of all, First of all the Structure of fuzzy neural network is constructed for classification of various texts present on web. Also, the Levenberg-Marquart optimizing algorithm is adopted for training of fuzzy neural network, and it

is shown that convergence rate and the classification accuracy of text classification is improved.

This paper has proven that the advantages of Fuzzy Theory and Neural Networks can be seen in Fuzzy Neural Networks and hence the Fuzzy Neural networks have better Self-adapting and approaching learning ability and also the better processing of Fuzzy information is possible in Fuzzy Neural Networks in comparison with BPNN networks

PROPOSED METHODOLOGY:

The Traditional BPNN plays an imperative role in the neural networks as a tool to resolve wide kinds of problems These feed forward nets consist of at least three layers (one input, one output, and at least one hidden layer) and use back propagation as learning mechanism.

The training of a network by back propagation involves several stages: calculation of input and output value, calculation of activation function, target function and back propagation of the associated error, the adjustment of the weight and the biases.

Semantic vector space generation with the help of Latent Semantic Indexing Technique.

Latent Semantic Indexing (LSI) is a technique that projects the original high dimensional document vectors into a space with "latent" semantic dimensions. Once a term by document matrix is constructed, LSA requires the SVD of this matrix to construct a semantic vector space which can be used to represent conceptual term-document associations.

LSI is an approach for automatic indexing and retrieval of documents based on the notion of concept. Founded on the vector space model (VSM), it was designed to overcome fundamental challenges such as retrieving documents based on the conceptual content, where individual words would provide unreliable evidence about the conceptual topic or meaning of a document.

Singular value decomposition

SVD is a well-developed method for extracting dominant features of large data sets and for reducing the dimensionality of the data. First of all a term document matrix can be constructed. Once a term by document matrix is constructed, SVD is used to decompose the term by document matrix to construct a semantic vector space which can be used to represent conceptual term-document associations.

- Term-document matrix is $m \times n$ matrix where m is number of terms and n is number of documents
- row of term-document matrix = term
- column of term-document matrix = document

$$A = \begin{bmatrix} d_1 & d_2 & \dots & d_n \\ \downarrow & \downarrow & & \downarrow \\ a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \begin{matrix} \leftarrow t_1 \\ \leftarrow t_2 \\ \vdots \\ \leftarrow t_m \end{matrix}$$

For every $m \times n$ matrix A there is singular value decomposition (SVD)

$$A = U \Sigma V^T$$

U orthogonal $m \times m$ matrix whose columns are left singular vectors of A

Σ diagonal matrix on whose diagonal are singular values of matrix A in descending order

V orthogonal $n \times n$ matrix whose columns are right singular vectors of A.

For LSI truncated SVD is used

$$A_k = U_k \Sigma_k V_k^T$$

Where,

U_k is $m \times k$ matrix whose columns are first k left singular vectors of A

Σ_k is $k \times k$ diagonal matrix whose diagonal is formed by k leading singular values of A

V_k is $n \times k$ matrix whose columns are first k right singular vectors of A

Rows of U_k = terms

Rows of V_k = documents

RESULT ANALYSIS:

To perform the experiment of our proposed methodology we use three categories of 20Newsgroup datasets i.e. sports document dataset, C.S document dataset and medicine document dataset. Each of the datasets contains 100 documents in which we use 60 documents for training and 40 documents for testing.

Various results were computed by applying only BPNN algorithm firstly and Then Combinatorial method of BPNN and LSI was applied. Results were computed based on precision, recall and f-measure to check the accuracy of the algorithm. The performance of our categorization system is evaluated by precision, recall and F-measure.

$$P = \frac{\text{No of correctly classified documents}}{\text{No of correctly retrieved documents}}$$

$$R = \frac{\text{No of correctly classified documents}}{\text{Total No of Relevant documents}}$$

$$F\text{-MEASURE} = \frac{2 * P * R}{P + R}$$

Dataset	Precision	Recall	F-measure
sports	0.9563	0.6843	0.797752152
CS	0.9362	0.7348	0.823362968
medicine	0.9175	0.6952	0.79102871

Table 1. Result Analysis of proposed work

Cluster	Precision	Recall	F-measure
0	0.8453	0.6843	0.75632687
1	0.8563	0.7354	0.791258428
2	0.8452	0.7653	0.803268004
3	0.9653	0.6943	0.807673885
4	0.9654	0.7431	0.839787814
5	0.9874	0.6854	0.80913912

Table 2. Result Analysis of Proposed work on Sports Dataset

Cluster	Precision	Recall	F-measure
0	0.9254	0.6523	0.7652132
1	0.9385	0.6754	0.7855046
2	0.9472	0.6843	0.7945681
3	0.9513	0.6932	0.8019959
4	0.9643	0.7154	0.8214088
5	0.9853	0.7254	0.8356072

Table 3. Result Analysis of Proposed Work on CS Dataset

Cluster	Precision	Recall	F-measure
0	0.8953	0.6743	0.769241577
1	0.8974	0.6831	0.775721531
2	0.9353	0.6943	0.796979369
3	0.9574	0.7104	0.815609737
4	0.9683	0.7357	0.836124777
5	0.9854	0.7467	0.849602425

Table 4. Result Analysis of Proposed Work on Medicine Dataset

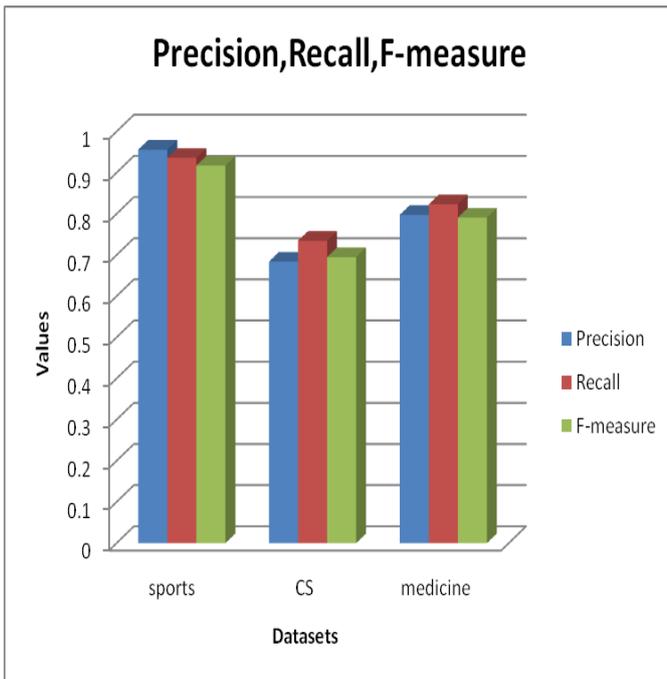


Figure 1. Comparison of Precision, recall and F-measure on different dataset for proposed work

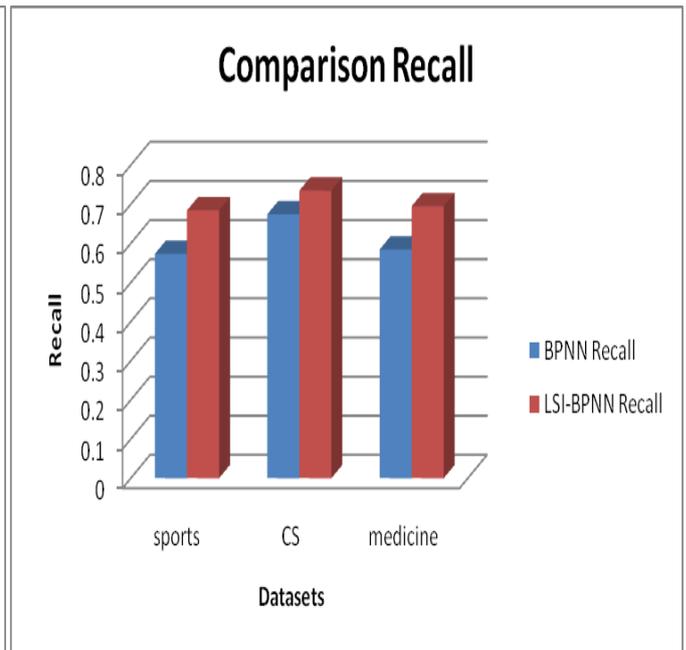


Figure 3. Comparison of recall on different dataset for proposed work for BPNN and LSI-BPNN

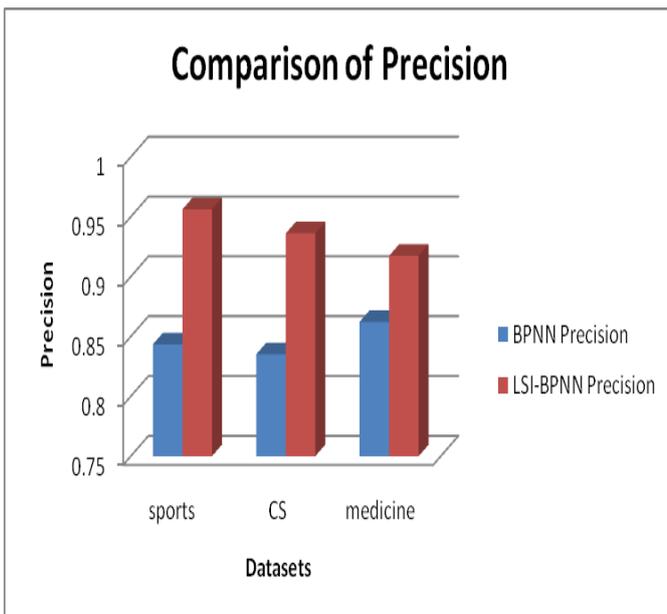


Figure 2. Comparison of Precision on different dataset for proposed work for BPNN and LSI-BPNN

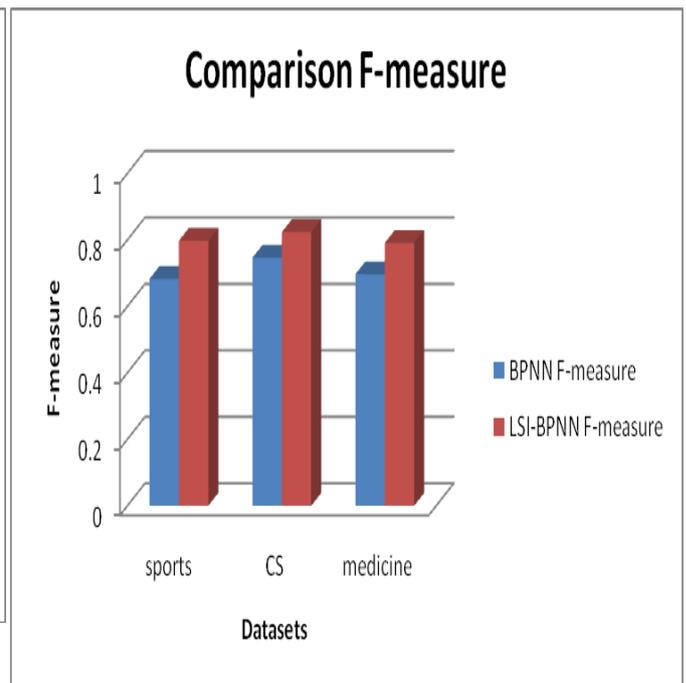


Figure 4. Comparison of F-measure on different dataset for proposed work for BPNN and LSI-BPNN

CONCLUSION:

The main aim is to test the classification accuracy on a set of documents based on the combinatorial method of LSI with BPNN. The results are then compared with the existing BPNN algorithm which is also used for the classification of text documents. The introduction of LSI reduces the dimension significantly and overcomes the troubles present in commonly used vector space model method for text depiction. The result analysis shows the performance of the proposed methodology.

Precision, Recall and F-measure of the proposed method is efficient as compared to the BPNN algorithm.

REFERENCES:

- [1] Zenglin Xu, Rong Jin, Kaizhu Huang, Michael R. Lyu, Irwin King “Semi-supervised Text Categorization by Active Search”, Proceedings of the 17th ACM conference on Information and knowledge management, pp. 1517-1518, 2008.
- [2] Robert E Uhrig “introduction to artificial neural network”, Proceedings of the IEEE 21st International Conference on Industrial Electronics, Control, and Instrumentation (IECON -1995), vol. 1, pp. 33-37, 1995.
- [3] Richard P Lippmann “Pattern Classification Using Neural Networks”, IEEE Communications Magazine, Volume 27, Issue: 11, pp. 47 – 50, 1989.
- [4] Mikhail Krivenko, Vitaly Vasilyev “ Sequential Latent Semantic Indexing” ,proceedings of the 2nd Workshop on Data Mining using Matrices and Tensors ,Article no3 , 2009
- [5] Berry M. W., Dumais S. T., O'Brien G. W. “Using linear algebra for intelligent information retrieval” // SIAM Review, Vol. 37. No. 4, pp-573–595, 1995
- [6] Jennifer Farkas “Documents, Concepts and Neural Networks” proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing - Volume 2, pp- 1021-1031
- [7] Ram Dayal Goyal “Knowledge Based Neural Network for Text Classification” GRC '07 Proceedings of the 2007 IEEE International Conference on Granular Computing, pp- 542-547, 2007
- [8] Zhihang Chen, Chengwen Ni and Yi L. Murphey “Neural Network Approaches for Text Document Categorization” IEEE 2006 International Joint Conference on Neural Networks,pp-1054-1060, 2006
- [9] Zhan Wang, Yifan He and Minghu Jiang “A Comparison among Three Neural Networks for Text Classification” published in Signal Processing, 2006 8th International Conference on (Volume:3), pp-102-107 ,2006
- [10] Cheng Hua Li and Soon Cheol Park “Neural Network for Text Classification Based on Singular Value Decomposition”, IEEE Computer and Information Technology, 2007. CIT 2007. 7th IEEE International Conference, pp.47-52, 2007
- [11] Tim Andersen, Wei Zhang “Features for Neural Net Based Region Identification of Newspaper Documents” IEEE Published in Document Analysis and Recognition,vol-1 2003. Proceedings. Seventh International Conference, pp- 403-407, 2003
- [12] Suvad Selman, Alma Husagic-Selman “Multilayered Feedforward Neural Networks as aTool for Distinction of the Authors of Texts” Published in Information, Communication and Automation Technologies (ICAT), 2011 XXIII International Symposium on, pp-1-6, 2011
- [13] Qiuming Ma, Zhiguang Qin, Fengli Zhang, Qiao Liu “Text Spam Neural Network Classification Algorithm” published in:Communications, Circuits and Systems (ICCCAS), 2010 International Conference on,pp-466-469, 2010
- [14] Diganta Saha “Web Text Classification Using a Neural Network” 2011 Second International Conference on Emerging Applications of Information Technology,pp-57-60, 2011
- [15] Jingwen Tian, Meijuan Gao, and Yang Sun “Study on Web Classification Mining Method Based on Fuzzy Neural Network” Published in Automation and Logistics, 2009. ICAL '09. IEEE International Conference on pp-1781 – 178, 2009