Volume 4, Issue 10, October 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Feature Selection for Plant Leaf Classification Based on Information Gain

C. S. Sumathi*
Assistant Professor, School of Post Graduate Studies,
Tamil Nadu Agricultural University,
Coimbatore-641003,
Tamil Nadu, India

A. V. Senthil Kumar
Director, Department of PG &
Research in Computer Applications,
Hindusthan College of Arts and Science,
Coimbatore-641028, Tamil Nadu, India

Abstract— Feature selection methods have been explored in the literature for the classification techniques, among which correlated feature, information gain, mutual information and chi-square are considered more effective. The leaf images contain inherent noise due to imaging equipment, operating environment and position of the image during image acquisition. In this paper, a method for classification of leaf images is proposed by exploiting the concept of information gain and explores the efficacy of learning algorithms of Multi-Layer Perceptron (MLP) for classifying plant leaf.This research shows information gain method for MLP with Batch Back propagation algorithm based learning increases computational efficiency by improving classification accuracy.It is observed that the proposed measure outperforms MLP with incremental training and Levenberg Marquardt based learning for plant leaf classification when tested with 9 species. Evaluation illustrates that information gain helps select features that result in significant improvements on MLP with Batch Back propagation algorithm classifier performance with an accuracy of 94.81%.

Keywords—Feature selection, Plant Leaf Classification, Information Gain, Multi-Layer Perceptron (MLP)

I. INTRODUCTION

Plants are primary among living organisms due to their ability to manufacture own food from simple inorganic materials and hence are considered Natural Resource Managers. So, we must understand what we manage and plant identification is a major component in such understanding. Biological scientists globally are trying to discover, describe and classify species. Due to many species extinction it is time to start a plant protection database. With information technology advances, there has been a explosive growth in abilities to generate and collect data in the last ten years. Very large commercial transaction databases were generated by retailers in business and voluminous scientific data was generated in varied scientific fields. The World Wide Web is an example with billions of web pages of textual and multimedia information used by millions. Analyzing voluminous data to be understood and used efficiently is a challenge. Data mining offsets this by ensuring techniques and software to automate analysis and exploration of huge complicated data sets. Data mining research was pursued by researchers in a various fields including statistics, machine learning, database management and data visualization [1]. Artificial Neural Networks (ANN) consists of many interconnected processing elements linked by weighted connections inspired by biological neurons. Learning in a biological system is through training or exposure to an input/output data set where a training algorithm adjusts weights iteratively. ANN are good pattern recognition engines and robust classifiers deciding about imprecise input data (Master, 1993) trained using Back Propagation Algorithm. Figure 1 reveals a block diagram representing an ANN.

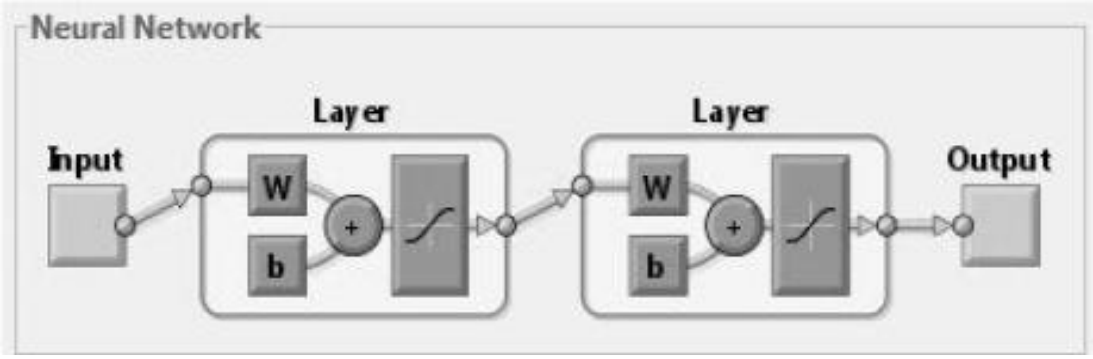

Fig.1 Representation of an Artificial Neural Network

Relative experimental studies have consistently shown Information Gain based feature selection to result in good classifier performance [2]. The principal task of feature selection applications is to improve the performance criterion such as accuracy.

II. RELATED WORKS

A compact weighted class association rule mining reflecting semantic significance of items considering its weight [3]. Classification constructs classifier and predicts new data instance. This new associative classification algorithm selects one non class informative attribute from dataset and all weighted class association rules are generated on that attribute's basis. The item's weight is considered as a parameter to generate weighted class association rules. The proposed algorithm calculates weight using HITS model. Experiments show that the new system generates less high quality rules thereby improving classification accuracy. Reference [4] analyzed medical image classification and retrieval systems have been finding extensive use in the areas of image classification according to imaging modalities, body part and diseases. One of the major challenges in the medical classification is the large size images leading to a large number of extracted features which is a burden for the classification algorithm and the resources. They proposed to investigate the efficacy of information gain of the extracted energy with respect to the class. Results obtained from information gain for global feature reduction method indicate the classification accuracy is not affected. Discernibility matrix method [5] described the best discriminate features using discernibility matrix and information gain is presented. The selection method using shows better results in terms of number of features selected and accuracy than applying methods individually.

Reference [6] proposed a new term weighting method for summarizing documents retrieved by IR systems. Unlike query biased summarization, this method does not use query information, but similarity information among original documents by hierarchical clustering and maps clusters similarity structure into weight of the every word, information gain ratio of probabilistic distribution of each word as term weight is adopted. A new information theoretic divisive algorithm for feature/word clustering and applying it to text classification was proposed [7]. Current techniques for words "distributional clustering" are agglomerative in nature resulting in (i) sub-optimal word clusters and (ii) high computational cost. To explicitly capture word clusters optimality in an information theoretic framework, a global criterion for feature clustering is first derived and then a fast, divisive algorithm that monotonically decreases the objective function value is presented. The algorithm minimizes "within-cluster Jensen-Shannon divergence" while simultaneously maximizing "between-cluster Jensen-Shannon divergence". In comparison to the earlier proposed agglomerative strategies, this divisive algorithm is faster and achieves comparable or higher classification accuracy. It also showed that feature clustering was effective to build smaller class models in hierarchical classification and presented detailed experiments using Naive Bayes and Support Vector Machines (SVM) on 20 Newsgroups data set and a HTML 3-level hierarchy documents collected from Open Directory project (www.dmoz.org).

Reference [8] stated that text classification was an important and studied area of pattern recognition, with various modern applications. Effective spam email filtering systems, automated document organization and management and improved information retrieval systems benefit from such techniques. The problem of feature selection or choosing relevant features from what is an incredibly large set of data is important for accurate text classification. It also provides a overview of text classification, followed by a survey of many feature selection methods used for text classification. Pruning techniques are briefly discussed to further reduce possible features (typically words) sets in a document before applying a feature selection method. A comparison of behavior of two most popular split functions, namely Gini Index function and Information Gain function was presented [9]. A situation where two split functions agree/disagree on selected split were mathematically characterized. Based on such characterizations it can analyze the frequency of agreement/disagreement of Gini Index function and Information Gain functions. It was seen that they disagreed only in 2% of cases, which explains why earlier published empirical results concluded that it was impossible to decide which of the two tests performed better. It also emphasized that methodology introduced in this paper is not limited to both analyzed split criteria using it to successfully formalize and compare other split criteria. Reference [10] discussed that the gain-based separation is a novel heuristic that modifies the standard multiclass decision tree learning algorithm to produce forests that can describe an example or object with multiple classifications. When the information gain at a node would be higher if all examples of a particular classification were removed, those examples are reserved for another tree. In this way, the algorithm performs some automated separation of classes into categories; classes are mutually exclusive within trees but not across trees. The algorithm was tested on naive subjects' descriptions of objects to a robot, using YUV color space and basic size and distance features. The new method outperforms the common strategy of separating multi label problems into L binary outcome decision trees, and also outperforms RA*k*EL a recent method for producing random multi label forests.

III. MATERIALS AND METHODS

A. Feature Selection

Information gain is a well known feature selection method. It is a reasonable objective to use for selecting feature. Using information gain will help to reduce the noise which is due to irrelevant features for influencing classifier. Information gain (IG) measures amount of information in bits about class prediction, when the only information available is presence of a feature and corresponding class distribution [2]. Information gain measure selects test attribute at every node in the tree [11]. The attribute with highest information gain (greatest entropy reduction) is selected as test attribute for current node. This attribute minimizes information needed to classify samples in resulting partitions. Entropy, measures amount of disorder or uncertainty in systems. In classification setting, higher entropy (more disorder) corresponds to sample of mixed label collection. Lower entropy corresponds to a case where there are pure partitions. In information theory, sample D's entropy is defined as follows:

$$H(D) = -\sum_{i=1}^{k} P(c_i \mid D) \log_2 P(c_i \mid D)$$

where $P(c_i \mid D)$ is probability of a data point in D being labeled with class c_i , and k is number of classes. $P(c_i \mid D)$ is estimated directly from the data as follows:

$$P(c_i \mid D) = \frac{|\{x_j \in D \mid x_j \text{ has label } y_j = c_i\}|}{|D|}$$

Also weighted entropy of a decision/split are defined as follows:

$$H(D_L, D_R) = \frac{|D_L|}{|D|} H(D_L) + \frac{|D_R|}{|D|} H(D_R)$$

where D was partitioned into D_L and D_R due to split decision. Finally, , information gain for a given split is defined as::

$$Gain(D, D_L, D_R) = H(D) - H(D_L, D_R)$$

In other words, Gain is anticipated entropy reduction caused by knowing an attribute's value.

B. Classification

1) Multi-Layer Perceptron (MLP) with various learning algorithms:

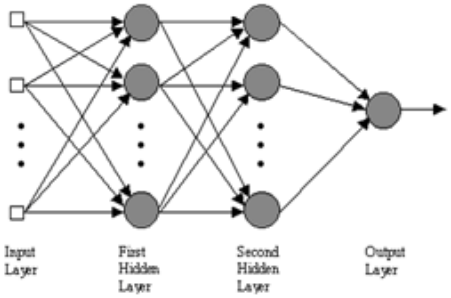

Fig. 2 Schematic representation of a multilayer perceptron

Neural Networks (NN) are an information processing technique based on how a biological nervous system, like the brain, processes information. The most common NN model is Multi-Layer Perceptron (MLP) (Fig. 2). This neural network is known as a supervised network as it needs a desired output to learn. The goal of this network is creating a model that correctly maps input to output using chronological data to ensure the model produces desired unknown output [12]. Figure1 shows MLP's graphical representation. The MLP and many other NN learn through an algorithm called backpropagation with which input data is repeatedly presented to NN. With every presentation, NN output is compared to desired output and an error computed. This error is fed back (backpropagated) to NN and adjusts weights so that error decreases with every iteration and neural model gets closer and closer to producing desired output. This is called training [13].

2) MLP with Levenberg–Marquardt based learning

Levenberg-Marquardt (LM) algorithm is a popular optimization algorithm which provides a numerical solution to function minimizing issues.Basically, it consists in solving the equation

$(JtJ + \lambda I)\delta = JtE$ (1)

Where **J** is Jacobian matrix for system, λ is Levenberg's damping factor, δ is weight update vector that is to be found and E is error vector with output errors for every input vector used on training network. The δ tell us by how much the network weights should be changed to achieve a (possibly) better solution.

JtJmatrix is also called the approximated Hessian. The λ damping factor is adjusted at every iteration, and guides optimization. When E reduction is rapid, a smaller value is used, bringing the algorithm closer to Gauss–Newton algorithm. If iteration gives insufficient reduction in residual, λ is increased, bringing it closer to gradient descent direction. Algorithm variations may include differing values for v, one for decreasing λ and the other to increase it [14].
Some advantages of LMA includes

1. The learning capability of the LMA is learned to be superior
2. LMA has rapid convergence advantages.
3. The LMA suits medium size datasets and is the fastest among training algorithms.

3) MLP with Batch Backpropagation algorithm based learning

Patterns are presented to a network before learning in batch training. In almost all cases, several passes must be made through training data. In batch training protocol, all training patterns are presented first and corresponding weight updates summed up; only then are actual network weights updated. This process is iterated till a stopping criterion is met [15]. Patterns need not be selected randomly, as weights are updated only after patterns are presented. In contrast, every weight change during continuous training reduces error for that instance, but decrease or increases error on training set as a whole. Hence, batch training ensures more accuracy.

Batch learning proceeds as follows:

begin initialize nH, w, criterion q, h, r ¬ 0
do r ¬ r + 1 (increment epoch)
m ¬ 0 ; Dwji ¬ 0; Dwkj ¬ 0
do m ¬ m + 1
xm¬select pattern
Dwji¬ Dwji + $\eta x_i \delta_j$; Dwkj¬ Dwkj + $\eta y_j \delta_k$
until m=n
wji¬ wji +Dwji; wkj¬ wkj +Dwkj
until $\|\nabla J(\mathbf{w})\| < \theta$
return w
end

4) MLP with Incremental Backpropagation algorithm based learning

The incremental back propagation updates the weights iteratively after each process [15]. After presentation of a training vector and computing the weight changes, they are introduced immediately to adapt the weights, without accumulating all changes until the end of the epoch. This sometimes referred as stochastic training. This has the ability to escape from entrapment at poor local minima on the error surface. This can happen as the shape of the error landscape changes with each process. Because of this, the instantaneous error may become of such magnitude that it allows the algorithm to jump out of the current landscape basin. Theoretically however, it is impossible to precisely investigate this behavior. Hence, the performance of the incremental backpropagation is laborious to analyze.

IV. RESULTS AND DISCUSSION

Nine species of plant leaves were selected with 15 samples each species. Sample image of the plant leaves used is shown in Fig. 3. The information gain based features were extracted using MatLab and classified using MLP with various learning method *viz.*, Levenberg–Marquardt, Batch Backpropagation, Incremental Backpropagation.

Fig. 3 Sample image of plant leaves

The classification accuracy obtained is tabulated is given in Table 1. Table 2 tabulates the precision, recall and fMeasure for various algorithms.

Table I Classification Accuracy

Technique Used	Classification accuracy
MLP with Levenberg–Marquardt learning	91.85%
MLP with Incremental Backpropagation learning	93.3%
MLP with Batch Backpropagation learning	94.81%

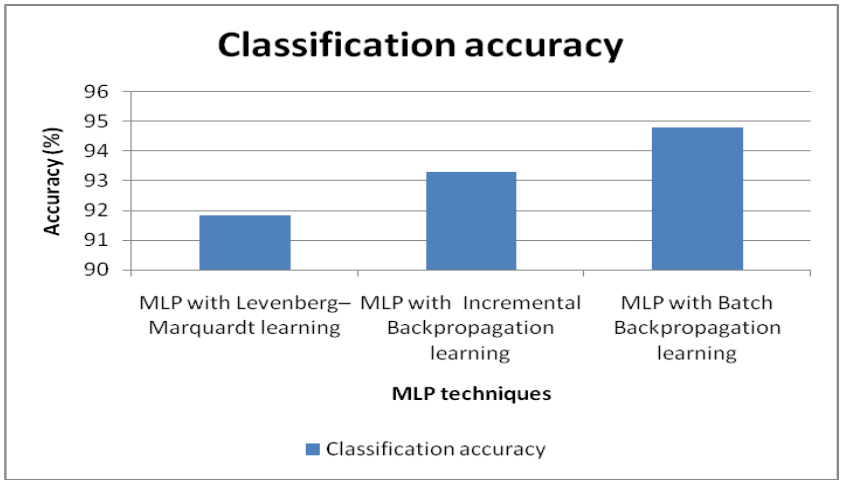

Fig. 4 Classification Accuracy

From Table 1 and Fig. 4 it is observed that the classification accuracy is achieved for different techniques. MLP with Batch Back Propagation learning achieves better accuracy as 3.22% than MLP with Levenberg-Marquardt learning and as 1.62% than MLP with Incremental Back Propagation Learning.

Table II Precision, recall and f Measure

Technique Used	Precision	Recall	f Measure
MLP with Levenberg–Marquardt learning	0.921	0.918	0.917
MLP with Incremental Back propagation learning	0.935	0.933	0.932
MLP with Batch Back propagation learning	0.949	0.948	0.947

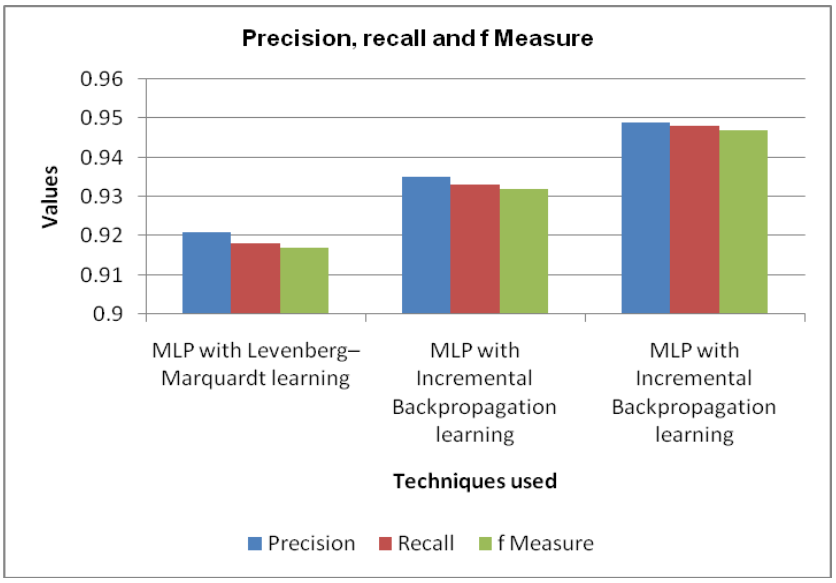

Fig. 5 Precision Recall and fmeasures

From Table 2 and Fig. 4 it is observed that the Precision Recall and fmeasures are achieved for different techniques. MLP with Batch Back Propagation learning achieves better precision as 3.04% than MLP with Levenberg-Marquardt learning and as 1.5% than MLP with Incremental Back Propagation Learning. MLP with Batch Back Propagation learning achieves better Recall as 3.27% than MLP with Levenberg-Marquardt learning and as 1.61% than MLP with Incremental Back Propagation Learning. MLP with Batch Back Propagation learning achieves better fmeasures as 3.27% than MLP with Levenberg-Marquardt learning and as 1.61% than MLP with Incremental Back Propagation Learning.

V. CONCLUSIONS

In this study, the information gain based features were extracted and classified using MLP with various learning method such as Levenberg–Marquardt, Batch Back propagation, Incremental Back propagation. Nine species of plant leaves were selected with 15 samples each species. Experimental results show that the proposed MLP with Batch Back Propagation learning method achieved a better performance of classification accuracy, precision, recall and fmeasures when compared to other MLP learning methods. The classification accuracy achieved by the proposed method is 94.9% which is the best accuracy when compared to other learning methods.

REFERENCES

[1] [Online]. Available: https://onlinecourses.science.psu.edu/stat557/node/1

[2] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons. 2012.

[3] S. P Ibrahim, and K. R. Chandran, "Compact Weighted Class Association Rule Mining using Information Gain," *arXiv preprint arXiv*, pp. *1112.2137*, 2011.

[4] T. Baranidharan, and D. K. Ghosh, "Medical Image Classification Using Information Gain for Global Feature Reduction," *database*, 11, 12.

[5] B.Azhagusundari, and A. S. Thanamani, Feature selection based on information gain. *International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN*, pp.2278-3075.

[6] T. Mori, "Information gain ratio as term weight: the case of summarization of ir results," in *Proc.* 19th international conference on Computational linguistics, Association for Computational Linguistics, August 2002, Volume 1, pp. 1-7.

[7] I. S Dhillon, S. Mallela, and R. Kumar, "A divisive information theoretic feature clustering algorithm for text classification," *The Journal of Machine Learning Research*, vol. 3, pp. 1265-1287, 2003.

[8] N.Nicolosi, *Feature Selection Methods for Text Classification*, 2008.

[9] L. E. Raileanu, and K. Stoffel, "Theoretical comparison between the gini index and information gain criteria," *Annals of Mathematics and Artificial Intelligence*, 41(1), pp.77-93, 2004.

[10] K. Gold, and A.Petrosino, 2010,"Using information gain to build meaningful decision forests for multilabel classification," in *Proc.*, IEEE 9th International Conference on Development and Learning (ICDL) August 2010, pp. 58-63.

[11] R. C.Barros, M. P. Basgalupp, A. C. de Carvalho, and A. A. Freitas, "Towards the automatic design of decision tree induction algorithms," in *Proc.*13th annual conference companion on Genetic and evolutionary computation (ACM), July 2011 pp. 567-574.

[12] H. C. Burger, C.J.Schuler, and S.Harmeling, "Image denoising with multi-layer perceptrons, part 1: comparison with existing algorithms and with bounds," *arXiv preprint arXiv:*1211.1544, 2012.

[13] D.T.Pham, and S.Sagiroglu,"Three methods of training multi-layer perceptrons to model a robot sensor," *Robotica*, 13(5), pp. 531-8,1995.

[14] M. I. Lourakis, "A brief description of the Levenberg-Marquardt algorithm implemented by levmar," Foundation of Research and Technology, vol.4, pp.1-6, 2005.

[15] L.Fu, H. H.Hsu, and J.C. Principe, "Incremental backpropagation learning networks. Neural Networks," IEEE Transactions 7(3),pp. 757-761, 1996.

