DECISION TREE LEARNING BASED FEATURE EVALUATION AND SELECTION FOR IMAGE CLASSIFICATION

HAN LIU[1], MIHAELA COCEA[1], WEILI DING[2]

[1]School of Computing, University of Portsmouth
Buckingham Building, Lion Terrace, Portsmouth PO1 3HE, United Kingdom
[2]Laboratory of Pattern Recognition and Intelligent Systems
Key Laboratory of Industrial Computer Control Engineering of Heibei Provience, Department of Automation
Institute of Electrical Engineering, Yanshan University, Qinghuangdao, 066004, China
E-MAIL: han.liu@port.ac.uk, mihaela.cocea@port.ac.uk, weiye51@ysu.edu.cn

Abstract:

In the big data era, machine learning has become an increasingly popular approach for data processing. Data could be in various forms, such as text, images, audios, videos and signals. The essence of machine learning is to learn any patterns from features of data. In the above types of data, the number of features is massively high, which could result in the presence of a large number of irrelevant features. Most machine learning algorithms are sensitive to irrelevant features so effective evaluation and selection of features in machine learning tasks are highly important. Also, effective evaluation of features can also help identify which features are necessary to be extracted from unstructured data. In this paper, we focus on the processing of image features in classification tasks. In particular, we review two main types of feature selection techniques, namely filter and wrapper. We also review several machine learning approaches that have been used popularly in image classification, and identify the limitations of these algorithms in terms of feature evaluation. An experimental study is reported showing the performance of C4.5 (a decision tree learning algorithm) and other popular algorithms (Naive Bayes, K Nearest Neighbours and Multi-layer Perceptron) on five image data sets from the UCI repository. Furthermore, we describe the nature of decision tree learning algorithms for analysing the capability of such algorithms in terms of feature evaluation in the training stage and for showing how rules extracted a decision tree can be used for evaluating features in the validation stage.

Keywords:

Data mining; Machine learning; Image classification; Decision tree learning; Feature evaluation

1. Introduction

Due to the vast and rapid increase in the size of data, machine learning has become a very powerful approach for processing of big data. Big data is generally characterised by five Vs: volume, velocity, variety, veracity and varability. The third V (variety) indicates that data can be in various forms: structured data and unstructured data. Unstructured data can be in the form of text, images, signals, audios and videos [1]. Such data needs to be transformed into structured data through feature extraction, such that traditional machine learning approaches can be used directly for learning from features of data. However, feature extraction from the above types of unstructured data usually results in a massively high dimensionality (the number of features), which is very likely to contain a large number of irrelevant features. The presence of irrelevant features not only increases the computational complexity of learning, but also leads to overfitting of training data, as most learning algorithms are sensitive to irrelevant features [2].

Due to the presence of irrelevant features, it has been highly important to effectively evaluate features and select only a subset of relevant features for learning models. In general, feature selection can be achieved in two approaches: filter and wrapper [3, 2, 4]. The filter approach aims to evaluate features prior to the training stage, which means to evaluate a set of features and select a subset of relevant ones for learning models. The wrapper approach aims to use an algorithm to learn models from different feature subsets and then check the predictive performance of the models for evaluating the corresponding subsets of features.

In this paper, we focus on investigating the capability of

some popular machine learning algorithms in terms of dealing with image features, since image features are among the most complex ones and thus effective evaluation and selection of features are highly important in image classification. Also, image classification has been involved in broad application areas. The simplest examples include handwritten digits recognition [5], which aims to identify a handwritten digit to be one of the 10 digits (0-9), and letter recognition [6], which aims to identify a letter to be one of the 26 letters (A-Z). In cancer research, classification of bio-medical images can help identify that a person is a patient of cancer or not [7]. In emotion recognition, image classification techniques can help identify the facial expression of people [8], such as sad, fear, anger and happy.

This paper is organized as follows: Section 2 reviews the two main approaches of feature selection (filter and wrapper), and analyses the capability of some popular machine learning approaches (e.g. decision tree learning, Naive Bayes and K Nearest Neighbours), in terms of evaluating image features. In Section 3, we conduct an experimental study by using five UCI data sets on image classification, and analyse the results in order to show that decision tree learning algorithms are highly competitive to those most popular algorithms used for image classification. Furthermore, we position in Section 4 the evaluation of image features in the context of decision tree learning, and explore how the metrics popularly used for evaluating the quality of decision trees can be used in different ways towards effectively evaluating image features. The contributions of this paper are highlighted in Section 5 and some further directions are suggested in this section towards advancing this area.

2. Related Work

In this section, we review two main types of feature selection techniques, namely filter and wrapper. We also review in general several popular machine learning algorithms, which include multi-layer perceptron, Naive Bayes, K nearest neighbours and C4.5 (an algorithm of decision tree learning). The review of these machine learning algorithms also involves their capability of evaluating image features.

2.1. Review of Feature Selection Techniques

As introduced in [9], feature selection involves four main steps: generation, evaluation, stopping criterion and validation. In particular, the generation procedure is aimed at generating a candidate feature subset. In the evaluation stage, a function is used to evaluate the subset of features selected in the generation stage. A stopping criteria is then used to decide whether it is necessary to stop the feature selection process. If yes, the selected feature subset is validated in the last stage. Otherwise, the feature selection process needs to be repeated through generation and evaluation of a candidate feature subset. The process of feature selection is illustrated in Fig. 1.

As mentioned in Section 1, feature selection techniques can be grouped into two categories, namely, filter and wrapper. The main difference between the two types of feature selections is in terms of the way of feature evaluation. The filter approach employs heuristics to rank the features according to their importance, whereas the wrapper approach employs an algorithm to learn classifiers from different subsets of features and then check the performance of these classifiers for evaluating the corresponding feature subsets. In terms of evaluation functions, popular heuristics employed by the filter approach include distance metrics [10], entropy [11], information gain [12], correlation coefficients [4], and covariance [13]. The wrapper approach just simply employs the error rate of a classifier as the evaluation function [9].

In terms of the performance of feature selection, the filter approach involves evaluation of features regardless of the fitness of the employed learning algorithm. In other words, a set of features is evaluated and the relevant ones are selected without considering that the selected feature subset is suitable or not for the chosen algorithm to learn a model. According to experimental results reported in [9], feature selection through the filter approach leads to a low level of time complexity. However, when the selected feature subset is used for a pre-employed algorithm to learn a model, the classification accuracy may be low due to the case that the feature subset is not suitable for the algorithm to do learning [2].

In contrast, the wrapper approach involves evaluation of features through checking the accuracy of the models learned from different subsets of features. In other words, a number (n) of different feature subsets are provided and an algorithm is used to learn n models from these feature subsets. The feature subset, which leads to the best model learned, is selected. According to experimental results reported in [9], feature selection through the wrapper approach leads to very high accuracy of classification but the time complexity is very high due to the case that all the possible combinations of features leading to different feature subsets need to be examined.

2.2. Review of Machine Learning Algorithms

In machine learning, the most popular algorithms include multi-layer perceptron (MLP), Naive Bayes (NB), K nearest neighbour (KNN) and C4.5. Also, the first four learning algorithms have been used popularly in image classification tasks.

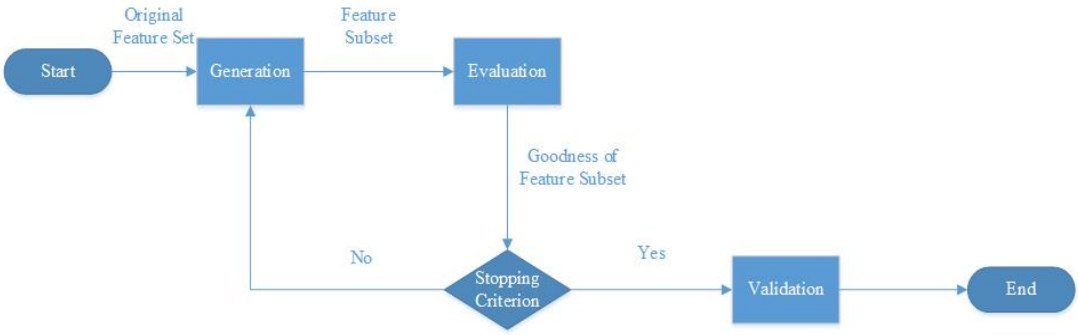

FIGURE 1. Feature Section Process

MLP is an algorithm of neural network learning. In a neural network topology, the first layer is referred to as the input layer and each node in this layer represents an input feature. The last layer is referred to as the output layer and the node in this layer represents the class output. There are also several other layers in the middle, which are referred to as hidden layers. In traditional machine learning, due to the constraints on hardware performance, there are typically only one or two hidden layers involved in the training stage. In the context of deep learning, the number of hidden layers is increased significantly towards advancing the performance of classifiers [14]. In the process of learning, the MLP algorithm works in a black box manner, which indicates that the mapping from inputs to outputs is poorly transparent. In terms of feature evaluation, the MLP algorithm can not show how features are evaluated and used in the training stage, nor can it achieve feature evaluation through examining the learned neural network model. More details on neural network learning can be found in [15].

NB is an algorithm of Bayesian learning. The NB algorithm employs the Bayes Theorem [16], which aims to learn the conditional relationships between feature values and class labels. When an instance is being classified, the classier learned by NB aims to judge the class to which has the highest conditional probability given the feature vector of this instance. The feature vector consists of a number (n) of values for the n features. In this context, the class, which has the highest conditional probability, would be assigned to the instance. In terms of feature evaluation, the NB algorithm can show how each input feature relates to a class in the context of probability theory. However, the nature of this algorithm assumes that all the input features are totally independent of each other, i.e. the NB algorithm does not involve identification of the correlation between different input features. More details on the NB algorithm and Bayesian learning can be found in [13].

KNN is an algorithm of lazy learning. The nature of this algorithm does not involve learning a model in the training stage. Instead, an instance is classified through finding k training instances that are the most similar ones to the unseen instance. In this context, the class to which the majority of the k instances belong would be assigned to the unseen instance. The similarity between two different instances can be measured through using distance functions, such as Euclidean distance [17]. In terms of feature evaluation, since the KNN algorithm does not involve learning a model, the features are not evaluated in the training stage. In the stage of classifying an unseen instance, the whole set of features are used equally, i.e. the features are not evaluated in terms of their importance or relevance. More details on the KNN algorithm and lazy learning can be found in [15].

C4.5 is an algorithm of decision tree learning. The nature of this algorithm involves the evaluation of features on an iterative basis. In particular, a decision tree has a root node and a number of internal nodes. Each of these nodes is labeled a feature and also known as a non-leaf node, which indicates that at each iteration of decision tree learning, a feature needs to be selected towards labelling a non-leaf node of a tree. The C4.5 algorithm employs entropy or information gain towards heuristic selection of features, which indicates that the selected feature is judged heuristically as the best one in the feature set. Following the completion of a learning task, as a decision tree works in a white box manner, the learned tree can show which features are appended into it as non-leaf nodes, i.e. those features, which are not appended into the tree, are judged as irrelevant according to the heuristic values. On the basis of the above statement, C4.5 can achieve self-evaluation of features in the process of learning a decision tree and after the completion of the learning. More details on the C4.5 algorithm and decision tree learning can be found in [15, 1]

TABLE 1. Data sets

Dataset	Feature Types	#Features	#Instances	#Classes
letter	Continuous	16	20000	26
optdigits	Continuous	64	5620	10
pendigits	Continuous	16	10992	10
segment	Continuous	19	2310	7
gisette	Continuous	5000	13500	2

TABLE 2. Classification accuracy on original data

Dataset	NB	MLP	KNN	C4.5
letter	64%	82%	95%	88%
optdigits	91%	98%	99%	91%
pendigits	86%	94%	99%	97%
segment	80%	96%	94%	97%
gisette	49%	93%	96%	94%

3. Experimental Studies

In this section, we conduct an experimental study by using five image data sets retrieved from the UCI repository [18]. The characteristics of these data sets are shown in Table 1.

In this experimental study, we compare the performance of several machine learning algorithms in terms of classification accuracy. These algorithms include MLP, NB, KNN and C4.5. We choose the first four algorithms as they have been popularly used in image classification tasks. C4.5 is not a popular method for image classification but we have identified in Section 2.2 this algorithm has advantages in terms of feature evaluation. Therefore, we compare C4.5 with those popular algorithms in terms of classification accuracy on image data, in order to show that the C4.5 algorithm has very similar performance to those popular algorithms but this algorithm also has a better capability of feature evaluation.

The experiments are conducted by partitioning a data set into a training set and a test set in the ratio of 70:30. On all the five data sets, the experiments are repeated 10 times and the average accuracy are used for comparison.

The results shown in Table 2 indicate that the C4.5 algorithm performs a very similar level of accuracy or even better than the other three algorithms, while the original data sets are used without pre-processing by feature selection techniques.

The results shown in Table 3 indicate that feature selection through the filter approach may help some algorithms build a better model but also negatively impact on the models learned by other algorithms. In particular, the five data sets are preprocessed for filtering irrelevant features by the Correlation-based Feature Subset Selection (CFSS) method [19]. The reduced feature set generally leads to margin decrease of classification accuracy for the MLP and C4.5 algorithms. For the NB algorithm, feature selection by the CFSS method results in an obvious increase of classification accuracy on the segment data set. For the KNN algorithm, feature selection by the CFSS method results in a margin increase of classification accuracy on the segment data set but also a margin decrease of performance on the letter data set.

TABLE 3. Classification accuracy on reduced data by feature selection

Dataset	NB	MLP	KNN	C4.5
letter	66%	78%	94%	87%
optdigits	91%	97%	99%	90%
pendigits	84%	94%	99%	96%
segment	87%	94%	95%	96%
gisette	49%	93%	96%	94%

The comparison of the results shown in Table 2 and Table 3 can indicate the point made in Section 2.1 that feature selection through the filter approach does not take into account the nature of learning algorithms and thus this way of feature selection may not necessarily leads to the increase of classification performance, i.e. the accuracy of classification may remain unchanged or even get lower, due to the case that the selected set of features does not fit well the nature of the chosen learning algorithms. As mentioned in Section 2.1, the wrapper approach has a very high level of time complexity although it may results in high classification accuracy. The above descriptions lead to the motivation of having learning algorithms capable of self-evaluation of features with lower time complexity.

4. Feature Evaluation by Decision Tree Learning

As discussed in Section 3, algorithms that are capable of self-evaluation of features in the process of learning are needed to minimize the computational complexity of feature selection. Also, as indicated in Section 3, C4.5 performs similarly or even better than the other popular learning algorithms in image classification. In this section, we position the evaluation of features in the context of decision tree learning. In particular, we show how features can be evaluated through using the decision tree models learned in the training stage.

In the context of decision tree learning, a tree can be converted directly into a set of rules, each of which is extracted from a branch of the tree. The set of rules can be evaluated in terms of their quality. Some popular measures of rule quality

include confidence and J-measure.

Confidence is a measure of the weight (predictive accuracy) of a rule, which is defined in Eq. 1. $P(x, y)$ is the joint probability that the antecedent and consequent of a rule both occur and $P(x)$ is the probability that the rule antecedent occurs independently.

$$Confidence = \frac{P(x, y)}{P(x)} \tag{1}$$

J-measure is a measure of the average information content of a rule, which is essentially the product of two terms as defined in Eq. 2. J-measure is also one of the most popular way of ranking rules [20].

$$J(Y, X = x) = P(x) \cdot j(Y, X = x) \tag{2}$$

The first term is the probability that the antecedent (left hand side) of a rule occurs and considered as a measure of simplicity [21]. The second term is read as j-measure, which is a measure of goodness-of-fit of a rule and also known as cross entropy [21]. The j-measure is defined in Eq. 3.

$$\begin{aligned} j(Y, X = x) = P(y|x) \cdot \log_2 \frac{P(y|x)}{P(x)} + \\ (1 - (P(y|x)) \cdot \log_2 \frac{1 - P(y|x)}{1 - P(x)} \end{aligned} \tag{3}$$

As mentioned above, a decision tree can be converted into a set of rules. These rules can be represented in the form of a rule based network [22]. The network topology is illustrated in 2.

In the first layer, each node represents an input feature. Each node in the middle layer represents a rule and the node in the last layer represents the class derived through using the rules in the middle layer. In terms of feature evaluation, this network topology can be used to interpret the ranking of features. In particular, the nodes in the input layer can be ordered to show the importance of features, e.g. the top one represents the most important feature and the bottom one represents the least important feature. Also, the connections between the nodes in the input layer and the nodes in the conjunction layer show which features are involved in which rules. An extreme case is that some nodes in the first layer do not have any connections to the nodes in the middle layer, which indicates that these features are not involved in any rules and thus considered as irrelevant.

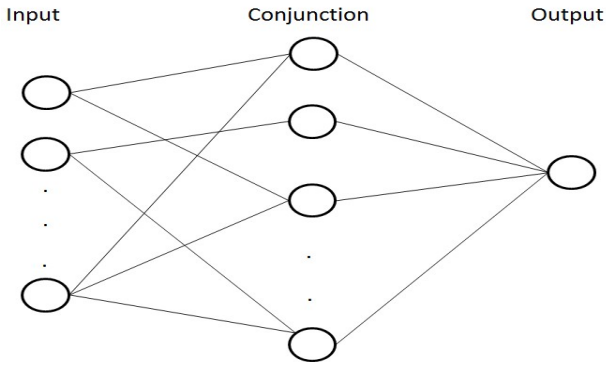

FIGURE 2. Rule Based Network Topology [22]

On the other hand, the nodes in the middle layer can also be ordered to show the ranking of these rules. In this context, we propose to evaluate features quantitatively according to the ranking and weight of rules. In particular, the ranking of rules can be achieved by using J-measure and the weight of each rule can be measured by using rule confidence. The evaluation function is defined in Eq. 4.

$$I(f) = \forall i(f \in r_i) \sum w(r_i) \cdot R(r_i) \tag{4}$$

$I(f)$ represents the importance of feature f, and r_i represents a rule, where i is the index of the rule. $R(r_i)$ is the rank of the rule r_i and $w(r_i)$ is the weight of the rule r_i. In addition, $\forall i(f \in r_i)$ means for all rules in which the feature f is involved, i.e. each rule covers a set of selected features.

5. Conclusions

In this paper, we compared C4.5 with three other algorithms (MLP, NB and KNN) in terms of classification accuracy on image data, The results show that the C4.5 algorithm performs similarly or even better than the other three ones. Based on these results, we positioned the study of image feature evaluation in the context of decision tree learning, as C4.5 has been judged more capable of evaluating features than other algorithms due to the nature of decision tree learning. In particular, decision tree learning algorithms can achieve self-evaluation of features in the process of learning, without the need to employ filter methods to pre-process features. Also, following the completion of a decision tree learning task, the learned tree can show explicitly which features have been used to label non-leaf nodes, i.e. the used features are considered as relevant. Moreover, we proposed to evaluate features through using the metrics of rule quality measure, i.e. the importance of features is highly correlated to the quality of rules in which the features are involved. In future, we will investigate empirically the performance of image classification while features are evaluated according to Eq. 4 in the context of decision tree learning.

Acknowledgements

This paper is supported by the Computational Intelligence Research Group in the School of Computing at the University of Portsmouth and by the China Scholarship Council.

References

[1] H. Liu, A. Gegov, and M. Cocea, *Rule Based Systems for Big Data: A Machine Learning Approach*. Switzerland: Springer, 2016.

[2] I. Guyon, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[3] P. Langley, "Selection of relevant features in machine learning," in *Proceedings of the AAAI Fall Symposium on Relevance*. Washington, D.C., USA: AAAI Press, 1994, pp. 127–131.

[4] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, Washington, D.C., USA, 21-24 August 2003, pp. 856–863.

[5] M. Buscema, "Metanet: The theory of independent judges," *Substance Use and Misuse*, vol. 33, no. 2, pp. 439–461, 1998.

[6] P. W. Frey and D. J. Slate, "Letter recognition using holland-style adaptive classifiers," *Machine Learning*, vol. 6, no. 2, pp. 161–182, 1991.

[7] R. S. Savage and Y. Yuan, "Predicting chemoinsensitivity in breast cancer with omics/digital pathology data fusion." *Royal Society Open Science*, vol. 3, no. 2, pp. 1–13, 2016.

[8] Y. Wang and H. Yu, "Facial expression-aware face frontalization," in *LNCS Proceedings of Asian Conference on Computer Vision*, Taibei, Taiwan, 20-24 November 2016, pp. 375–388.

[9] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, pp. 131–156, 1997.

[10] P. Montalto, M. Aliotta, A. Cannata, C. Cassisi, and A. Pulvirenti, "Similarity measures and dimensionality reduction techniques for time series data mining," in *Advances in Data Mining Knowledge Discovery and Applications*, A. Karahoca, Ed. InTech, 2012, pp. 71–96.

[11] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[12] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[13] D. Barber, *Bayesian Reasoning and Machine Learning*. Cambridge: Cambridge University Press, 2012.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge: MIT Press, 2016.

[15] T. Mitchell, *Machine Learning*. London: McGraw Hill,, 1997.

[16] M. Hazewinkel, *Bayes formula*. London: Springer,, 2001.

[17] H. Liu, A. Gegov, and M. Cocea, "Nature and biology inspired approach of classification towards reduction of bias in machine learning," in *International Conference on Machine Learning and Cybernetics*, Jeju Island, South Korea, 10-13 July 2016, pp. 588–593.

[18] M. Lichman, "UCI machine learning repository, http://archive.ics.uci.edu/ml," 2013.

[19] M. A. Hall and L. A. Smith, "Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper," in *Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference*, Orlando, Florida, 1-5 May 1999, pp. 235–239.

[20] P.-N. Tan, V. Kumar, and J. Srivastava, "Selecting the right objective measure for association analysis," *Information Systems - Knowledge discovery and data mining*, vol. 29, no. 4, pp. 293–313, 2004.

[21] P. Symth and R. M. Goodman, "An information theoretic approach to rule induction from databases," *Royal Society Open Science*, vol. 4, no. 4, pp. 301–316, 1992.

[22] H. Liu, A. Gegov, and M. Cocea, "Rule based networks: An efficient and interpretable representation of computational models," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 7, no. 2, pp. 111–123, 2017.