Journal of Engineering and Applied Sciences 12 (7): 1890-1895, 2017
ISSN: 1816-949X

Presentation and Evaluation of an Effective Algorithm for Clustering

R. Alesheykh
Faculty of Engineering, Payame Noor University (PNU), P.O. Box 19395-3697 Tehran, IR of Iran

Abstract: Clustering algorithms partition a set of data into numbers of groups according to their similarity. A clustering algorithm is a common technique for statistical data analysis and used in many fields including information retrieval and machine learning. In the current research some machine learning algorithms have been presented to identify three timber species and group them into the correct clusters. Machine learning algorithms such as C4.5 decision tree, RIPPER rule learning method and bayesian network have been experimented across Winger-Ville distribution method to do the identification task. The employment of the most suitable timber for each specific purpose demands for the development of an effective computerized method for the identification of timber species. Since, each species creates different properties in timber, a reliable and powerful evaluation approach of identification plays an important role in suitably use of the timber. The final analysis shows the clustering performance of 91% when the output of Winger-Ville distribution method is employed by RIPPER rule learning algorithm.

Key words: Machine learning, C4.5 decision tree, RIPPER rule learning, bayesian networks, Winger-Ville distribution

INTRODUCTION

Identifying the type of timber can be a challenging task because of the variables that combine to give timber its appearance. Many timber species can be stained to look like other timbers and may not be simply distinguishable. It is important that professionals be able to distinguish the timber of one species from another. Timber species verification is particularly important when musical instruments are made and original timbers should be used. The cost of timber utilized in furniture manufacturing also depends on types of the timber. Each kind of timber has its unique structure, physical and mechanical properties and is used for different application. The differences in timber properties allow for the manufacture of timber products with many different appearances and uses. Depending on the timber application it is essential for a production line to improve their timber processing from the earliest stage of rough milling in order to increase timber yield and maintain a reliable quality of product output (Hashim *et al*., 2015). Timber of a particular species is identified by its features including strength, density, hardness, image and sound produced from timber. Reliable timber identification usually requires the ability to recognize basic differences in timber and this will determine the suitability of the timber for a particular use. At present, the international feature detection on the surface of the timber uses non-contact detection method with more voice control, optical control and image recognition (Chen *et al*., 2014). Human vision is capable of recognizing the patterns of timber and distinguishes between patterns but describing the difference precisely is not easy (Hashim *et al*., 2016). Since, sound processing techniques have been found to be more accurate than the conventional task of visual inspection in assessing timber species in this research the focus has been laid exclusively on sound processing techniques.

The objective of this research is to examine some machine learning algorithms for clustering three kinds of timbers and to detect the correct timber species which is beech, alder or maple. The purpose is to assess the relative performance of some well-known machine learning algorithms and aid to increase the reliability and consistency of the clustering system. In this research, machine learning algorithms such as decision tree, rule learning method and bayesian network is tested across Winger-Ville Distribution signal analysis method. While the origins of these machine learning approaches are distinct and the underlying algorithms differ substantially, the fundamental process is the same; they are all inductive methods. The above mentioned algorithms were performed using WEKA http://sourceforge.net/projects/weka/software which provides a safe chance of testing several machine learning algorithms. The final goal of this study is to develop a computerized system for timber species determination.

Data preparation: For experimental task, 90 surfaced boards with approximately 1/2 inch thickness were collected from a carpenter. Thirty boards were chosen for each timber type. All measurements were made by a microphone and a conventional PC sound card which could sample in stereo. The signals recorded on each timber were listened carefully to evaluate the sound quality from each timber. The examination was done with a hammer which produced different sounds when hitting on different kinds of timbers. Due to the lack of knowledge obtained from sound inspection which only explains different sounds, use of machine learning approach seems efficient.

MATERIALS AND METHODS

Wigner ville distribution method: The aim of feature extraction is to present signals compactly and efficiently. Features are extracted to obtain the most significant information from the original data with an aim of reducing computational burden for further clustering task. In order to analyze a signal whose component frequencies vary in time, a time-frequency distribution of the signal is a safe choice. The time-frequency technique that is mentioned and employed in this work is Winger-Ville Distribution (WVD) method. The WVD method is a two-dimensional function describing the frequency content of a signal as a function of time (Quian and Chen, 1996) and possesses many advantageous properties. WVD is a very useful tool when one needs to analyze time-frequency representations of non-stationary signals. The WVD graphical appearance is very similar to a signal's spectrogram and the results of the common spectrogram can be compared with the WVD of the signal. In this study, signals were collected by making tests on 90 surfaced boards and were used in pre-processing and feature extraction stages. The output of feature extraction process together with Principle Component Analysis (PCA) is 40 features extracted by WVD method. The feature vector is then presented to the machine learning algorithms for further clustering task concerning the determination of timber species.

Machine learning algorithms: The field of machine learning is concerned with the question of how to construct computer programs that automatically improves with experience (Sugiyama, 2015, 2013). Given that, each machine learning method has its strengths and limitations and that real world problems do not always satisfy the assumptions of a particular method, one approach is to apply many appropriate methods and select the one that provides the best solution. This study explores the application of effective machine learning algorithms to overcome challenges associated with data analysis and demonstrates how machine learning algorithms and signal processing techniques have contributed and are contributing to the research (Yu *et al*., 2015). In this research, some machine learning algorithms such as decision tree, rule learning method and bayesian network are examined in WEKA Software and the results concluded from the mentioned methods will be compared. What follows next is a brief discussion concerning the above mentioned methods.

Decision tree algorithm: Decision tree is a predictive model that maps target values from observations. It is a flow-chart-like tree structure where each internal node denotes a test on an attribute, each branch represents an outcome of the test and leaf nodes represent classes or class distributions (Han *et al*., 2011; Ma *et al*., 2016). In order to classify an unknown sample using decision trees, the attribute values of the sample are tested against the decision tree. To learn which attribute should be tested at the root of the tree, each instance attribute is evaluated using a statistical test. This test is done to determine how well each attribute alone classifies the training examples. A path is traced from the root to a leaf node that holds the class prediction for the sample. The selection of attributes that separated the internal nodes is very important during the construction process and determines the final structure of the wide range of decision trees (Susanto, 2013). The popular algorithm which has been used for generating decision tree in the current work is C4.5 (Sugiyama, 2016). The C4.5 algorithm uses a divide-and-conquer approach for growing decision trees. The splitting node strategy is based on the computation of the information gain ratio. The basic idea is that each node should hold a question concerning the attribute which is the most informative among the set of attributes not yet considered in the path from the root to that node. Information value called entropy, also measures how informative is the association of an attribute with a node (Robert, 2014). The sub-trees are spanned by splitting the training dataset according to this strategy. Once the initial decision tree is constructed, a pruning procedure is initiated to decrease the overall tree size and decrease the estimated error rate of the tree (Quinlan, 1993).

Rule learner algorithm: Rule learner (rule induction) method applies an iterative process consisting in first generating a rule that covers a subset of the training examples and then removing all examples covered by the

rule from the training set before subsequent rules are learned. This process is repeated iteratively until there are no examples left to cover. The final rule set is the collection of the rules discovered at every iteration of the process. Rule learner algorithms expect positive and negative examples for an unknown concept. If any of the learned rules fires for a given example, the example is classified as positive and if no rule fires, it is classified as negative (Furnkranz, 1999). The rule learner algorithm employed in this work is Repeated Incremental Pruning to Produce Error Reduction (RIPPER). RIPPER builds a rule set by repeatedly adding rules to an empty rule set until all positive examples are covered. Rules are formed by greedily adding conditions to the antecedent of a rule (starting with empty antecedent) until no negative examples are covered. The pruning stage then attempts to simplify the rule by removing a sequence of conditions at the end of the rule. This greedy process examines which deleted sequence maximizes the proportion of positive examples over total examples covered. Afterward, a rule set is constructed an optimization post pass massages the rule set so as to reduce its size and improve its fit to the training data. The optimization stage examines each rule in sequence and decides whether the rule needs to be replaced, revised or kept.

Rule induction and decision tree methods both split a data set into subgroups on the basis of the relationships between predictors and the output field. Rules can be symmetric whereas trees must select one attribute to split on first and this can lead to trees that are much larger than an equivalent set of rules (Witten *et al*., 2011).

Bayesian network learning: Bayesian Networks (BNs) (Pearl, 1988; Rancoita *et al*., 2016; Fuster *et al*., 2016) are a probabilistic framework for reasoning under uncertainty. BNs are directed acyclic graphs where the nodes are random variables which denote attributes, features or hypothesis and the arcs specify the conditional independencies between the random variables. Associated with each node (child node) is a probability distribution on that node given the state of its parent nodes. A bayesian network specifies a joint distribution in a structured form. The joint distribution described by a graph is computed by the product of conditional probabilities for each node conditioned on the variables corresponding to the parents of that node in the following way:

$$P(y_1,\ldots,y_n)=\prod_{i=1}^{n}P(y_i \mid \text{Parents}(y_i))$$

Where:

y_i = The value of the random variable Y_i

Parent (Y_i) = The value of the parents of Y_i

In order to specify the probability distribution of a BN, one must give prior probabilities for all root nodes and conditional probabilities for all other nodes, given all possible combinations of their direct predecessors. Once the network is constructed, it constitutes an efficient device to perform probabilistic inference. Many algorithms have been proposed on learning Bayesian network structure. One method is score-and-search approach (Heckerman, 1996; Suzuki, 1999) which poses the learning problem as a structure optimization problem. Namely, it uses a score metric to evaluate every candidate network structure and then, finds a network structure with the best score. In the current work, the bayesian network represents the probabilistic relations between extracted features and the target class which is a timber species. Given the features, the network computes the probabilities of being kept in either beech, alder or maple cluster. BN learning algorithm also uses the general purpose search method of simulated annealing to find a well scoring network structure.

RESULTS AND DISCUSSION

In this research, machine learning algorithms such as C4.5 decision tree, RIPPER rule learning method and bayesian networks were tested for clustering the timber species into three clusters named beech, alder and maple. WEKA, an open source machine learning framework which is a collection of machine learning algorithms was employed to do the clustering task. Clustering, called unsupervised classification is the process of segmenting heterogeneous data objects into a number of homogenous clusters. Each cluster is a collection of data objects that are similar to one another and dissimilar to the data objects in other cluster (Kuzelewska, 2014; Ali *et al*., 2016). WVD feature extraction method together with PCA technique was used to extract the most important features of the signals obtained from timbers. Using PCA, the redundant features were removed effectively and this made the efficiency be improved.

Before proceeding any further with the clustering process, it is worth mentioning that the sound signals which collected by making experiments on 90 surfaced boards were partitioned into training and test sets. Since the clustering rate reported for the current work is based on the analysis of a very small set of data and to investigate how the discussing methods are performed on new or different data sets, cross-validation has been used.

Table 1: Results with different machine learning algorithm

Machine learning algorithms/Experiments	C4.5 (%)	RIPPER	Bayesian network (%)
WVD with 10-fold cross-validation	90	91	85
WVD with 5-fold cross-validation	88	90	80
WVD with training and testing sets	84	86	76

Cross-validation is a method for evaluating machine learning algorithm by dividing data into training and testing sets. In cross-validation a fixed number of partitions of the data called folds are determined. In this experiment 5 and10-fold cross-validation has been chosen for partitioning the dataset. This means that the data is split into five and ten approximately equal partitions and each in turn is used for testing and the remainder is used for training. The procedure is repeated five and ten times so that, by the end, every instance has been used exactly once for the testing. Many experiments on numerous datasets have shown that 10-fold cross validation is about the right number of folds to get more robust results as clustering rate.

The results obtained from each machine learning algorithm with the mentioned WVD method have been reported in table. Table 1 reports the percentage of clustering using 5-fold and 10-fold cross-validation. The last row of the table also shows clustering performance using partitioning the data into 80 and 20% for predetermined training and testing set respectively. Results from the experiments show that the proposed RIPPER rule learning algorithm together with WVD feature set has achieved superior clustering accuracy of 91% as compared to other machine learning algorithms. Moreover, the results demonstrate acceptable clustering accuracy across timber species meaning that the proposed features could be generalized to other timber species as well. The table also indicates that the entire machine learning algorithm employed in this work performed much better when using 10-fold cross-validation. Cross-validation is intended to avoid the possible bias introduced by relying on any one particular division into test and train components. By partitioning the original set into several parts and compute an average score over the different partitions, i.e., average number of corrected classified samples over all the samples in every partition, more reliable result will be concluded.

According to the results comprehended from the above tables, RIPPER rule learning algorithm has achieved better clustering accuracy in contrast with C4.5 decision tree. The reason might be that rules are much more compact than trees and a default rule can cover cases not specified by other rules (Witten *et al*., 2011). When a decision tree is built, many of its branches may reflect anomalies in training data. In addition, when adding new rules to an existing rule set, there is no need to disturb previous rules but to add a tree structure may require modifying the whole tree. The experiments also indicate that bayesian network shows lower performance as compared to two other machine learning algorithms. This is because bayesian network requires initial knowledge for assigning probabilities. Either an expert must provide prior probabilities for all root nodes and conditional probabilities for all other nodes or they can be obtained from an algorithm which automatically induces them. The quality of the results of the network strongly depends on the quality of the prior beliefs. The accuracies obtained in this study are in line with or higher than the accuracies reported in comparable studies. In research (Yusof *et al*., 2013) the use of fuzzy logic-based pre-classifier as a means of treating uncertainty to improve the classification accuracy of tropical timber recognition system was proposed. The pre-classifier serves as a clustering mechanism for the large database simplifying the classification process making it more efficient. The classification accuracy showed 88.9 % precisian without performing fuzzy logic pre-classifier and 93% precisian after performing fuzzy logic pre-classifier. In study (Lei and Yan, 2010) the recognition approach of timber species on the basis of mathematical simulation theory was proposed and a hexagon mathematics model for timber cells was established. Various parameters namely area, perimeter, roundness, the width of the diameter and the thickness of the diameter were extracted and obtained. This method greatly accelerates the speed of recognition and comparison and it has also reduced the uncertainty of the traditional method which depends mainly on image pixels characteristics. The recognition reliability of the results differs from 91.3-91.5 for fir samples in work (Lei and Yan, 2010).

CONCLUSION

In this study, some machine leaning algorithms were proposed for recognizing three timber species and group them into the correct clusters. In modern industry automating the identification of timber species is one of the key issues in increasing the product quality and a non-destructive method of recognition facilitates the task for users. The use of WVD method described in this study successfully differentiates between various kinds of timber samples. After several experiments, the final clustering rate demonstrated a precision of 91% using the

combination of RIPPER rule learning algorithm with WVD feature extraction method. Results from the analysis reveals that the proposed rule learning learning algorithm provides better performance than other machine learning methods and performs acceptably well across multiple timber species. The clustering rate shows that it is possible to detect and cluster timber species systematically and build an accurate automated inspection procedure.

The relative performance and clustering efficiency of the techniques used in the current case will become more apparent when they are applied to a larger database of timber species. Although, 91% clustering accuracy seems so reliable, it would be more efficient if one tries more other machine learning algorithms for the clustering task to explore the highest performance possible.

SUGGESTIONS

In the future, it is possible to extend the work further by inspecting more popular clustering algorithms to improve the class discrimination ability. Also performing additional pre-processing step after feature extraction phases is recommended to get more meaningful results. Another interesting aspect that could be worth investigating is to combine features extracted from sound processing techniques with the features extracted from image processing techniques and propose one single vector to the algorithms.

REFERENCES

Ali, S.H., A.I. El Desouky and A.I. Saleh, 2016. Notice of retraction: A new profile learning model for recommendation system based on machine learning technique. Indonesian J. Electr. Eng. Inf., 4: 81-92.

Chen, L., K. Wang, Y. Xie and J. Wang, 2014. The segmentation of timber defects based on color and the mathematical morphology. Optik Intl. J. Light Electron. Optics, 125: 965-967.

Furnkranz, J., 1999. Separate-and-conquer rule learning. Artif. Intell. Rev., 13: 3-54.

Fuster, P.P., P. Tauler, M.B. Veny, A. Ligeza and A.A.L. Gonzalez *et al*., 2016. Bayesian network modeling: A case study of an epidemiologic system analysis of cardiovascular risk. Comput. Methods Programs Biomed., 126: 128-142.

Han, J., M. Kamber and J. Pei, 2011. Data Mining: Concepts and Techniques. 3rd Edn., Morgan Kaufmann Publishers, USA., ISBN-13: 9780123814791, Pages: 744.

Hashim, U., S. Hashim and A. Muda, 2015. Image collection for non-segmenting approach of timber surface defect detection. Intl. J. adv. Soft Comput., 7: 15-34.

Hashim, U., S. Hashim and A. Muda, 2016. Performance evaluation of multivariate texture descriptor for classification of timber defect. Opt. Intl. J. Light Electron Opt., 127: 6071-6080.

Heckerman, D., 1996. A tutorial with learning Bayesian networks. Microsoft Research, Redmond, Washington. http://sfi.nr.no/sfi/images/8/8f/Heckermantutorial.pdf.

Kuzelewska, U., 2014. Clustering algorithms in hybrid recommender system on MovieLens data. Stud. Logic Grammar Rhetoric, 37: 125-139.

Lei, Z. and M. Yan, 2010. Timber species recognition approach based on mathematic simulation theory. Proceedings of the 2010 International Conference on Information Science and Management Engineering (ISME) Vol. 2, August 7-8, 2010, IEEE, Harbin, China, ISBN:978-1-4244-7669-5, pp: 197-202.

Ma, L., S. Destercke and Y. Wang, 2016. Online active learning of decision trees with evidential data. Pattern Recognit., 52: 33-45.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo, USA., ISBN: 9781558604797, Pages: 552.

Quian, S. and D. Chen, 1996. Joint Time Frequency Analysis: Methods and Applications. Prentice Hall, Upper Saddle River, New Jersey, ISBN: 9780132543842, Pages: 302.

Quinlan, J.R., 1993. Programs for Machine Learning. 1st Edn., Morgan Kaufmann, San Francisco, ISBN: 1-55860-238-0.

Rancoita, P.M., M. Zaffalon, E. Zucca, F. Bertoni and C.P. De Campos, 2016. Bayesian network data imputation with application to survival tree analysis. Comput. Stat. Data Anal., 93: 373-387.

Robert, M.G., 2014. Entropy and Information Theory. 2nd Edn., Springer, USA., ISBN:9781489981325, Pages: 409.

Sugiyama, M., 2015. Statistical Machine Learning. In: Introduction to Statistical Machine Learning, Sugiyama, M. (Ed.). MK Publishers, ?Burlington, Massachusetts, ISBN-13: 978-0128021217, pp: 3-8.

Susanto, B.M., 2013. Naive Bayes Decision Tree Hybrid Approach for Intrusion Detection System. In: Bulletin of Electrical Engineering and Informatics, Sutikno, T., J. Auzani and F. Moch (Eds.). Institute of Advanced Engineering and Science, Indonesia, Asia, ISBN: 0-201-14455-7, pp: 225-232.

Suzuki, J., 1999. Learning Bayesian belief networks based on the minimum description length principle: Basic properties. IEICE. Trans. Fundam. Electron. Commun. Comput. Sci., 82: 2237-2245.

Witten, I.H., E. Frank and A.H. Mark, 2011. Data Mining: Practical Machine Learning Tools and Techniques. 3rd Edn., Morgan Kaufmann, San Francisco, CA., USA.

Yu, J., H. Zhou and X. Gao, 2015. Machine learning and signal processing for human pose recovery and behavior analysis. Signal Process., 110: 1-4.

Yusof, R., M. Khalid and A.S.M. Khairuddin, 2013. Fuzzy logic-based pre-classifier for tropical wood species recognition system. Mach. Vision Appl., 24: 1589-1604.