Application of Weighted Classification of Non-equilibrium Decision Tree in Emotion Analysis of Poetry Reading

LI Feng-kun*, Zhang Yong

1 Dalian Neusoft University of Information, Dalian 116023, China
2 Dalian Neusoft University of Information, Dalian 116023, China
*Corresponding author’s e-mail: lifengkun@neusoft.edu.cn

Abstract. In this paper, a new unbalanced decision tree algorithm for appealing expressions of reading poem is proposed. This algorithm called Weighted Division of Unbalanced Decision Tree (WDOUDT). Compared to the traditional decision tree, it has fewer nodes and more interpretability, faster convergence and higher accuracy. WDOUDT is used to identify the moving expression when students read poem. The results show that WDOUDT performed better in accuracy than decision tree, and its time complexity is lower than that of the classical decision tree algorithm. Besides, it has good generalization ability and good robustness against noise data.

1. knowledge preparation

Relief_f algorithm, k-means, kmodes and KPrototype algorithm are needed for the decision tree derivation algorithm proposed in this paper. In order to illustrate the algorithms better in this paper, this chapter first briefly reviews these algorithms. So this chapter begins with a brief review of these algorithms..

1.1. The algorithms of relief and relief_f

Attributes are also called features, for a classification task, given a set of attributes, some attributes may be useful, called related characteristics and other properties are useless and are called irrelevant features. Relief is a very famous filtering feature selection algorithm which was proposed in 1992 by Kira K Rendell L A. The Relief algorithm only applies to dichotomous problems. In 1994, Relief_f was proposed by Kononenko, who extended the Relief algorithm to multi-classification problems in Literature [1]. The algorithm assigns different weights to attributes according to their relevance. The algorithm selects a sample R from training set D randomly, and then looks for the nearest neighbor sample H from samples similar to R, which is called Near Hit. Find the nearest neighbor sample M from samples that are different from R, called Near Miss, Then the weights for each feature are updated according to the following rules. If the distance between R and Near Hit on some feature is less than the distance between R and Near Miss, it indicates that the feature is beneficial to distinguish the nearest neighbors of the same and different species, then the weight of the feature is increased. Conversely, if the distance between R and Near Hit on a feature is greater than the distance between R and Near Miss, it indicates that the feature has a negative effect on the distinction between the nearest neighbors of the same and different species, then the weight of the feature is reduced. The above process was repeated m times, and finally the average weight of each feature was obtained. The greater the weight of the feature, the classification ability of the feature is stronger; otherwise, the classification ability of the feature is weaker.
\[ W(A) = W(A) - \frac{\text{diff}(A,R,H)}{m} + \frac{\text{diff}(A,R,M)}{m} \] (1)

In formula 1, \( W(A) \) is the weight of attribute A, R is the randomly selected sample, H is the nearest neighbor of R of correct guessing, M is the nearest neighbor of R of error guessing. Diff() is the function used to calculate the distance between two samples on attribute A. When A is a continuous attribute, diff is the Manhattan distance normalized to the interval \([0,1]\). When A is the discrete attribute, the attribute values of the two samples are the same, therefore \( \text{diff}=0 \), otherwise, \( \text{diff}=1 \), namely,

\[ \text{diff}(A,R_1,R_2) = \frac{|R_1(A)-R_2(A)|}{\max(A)-\min(A)} \] (2)

The Relief algorithm only applies to dichotomous problems. In 1994, Relief_f was proposed by Kononenko, who extended the Relief algorithm to multi-classification problems[2]. The weight is calculated according to the following formula 3:

\[ W(A) = W(A) - \frac{\text{diff}(A,R,H)}{m} + \sum_{C \in \text{class}(R)} P(C) \times \text{diff}(A,R,M|C) \] (3)

In the above formula, C represents other categories different from the category of sample R, and \( P(C) \) represents the proportion of samples of category C in all samples different from R. The running time of Relief and Relief_f algorithms increases linearly with the increase of the number of samples m and the number of original features N, therefore, the operation efficiency is very high, which can be used for large data sets. In this paper, Relief_f algorithm is used to weight the attributes. When the nodes are divided and calculated, the distance with weights is used instead of the actual distance.

1.2. Kmeans, KMode, KPrototype

Kmeans clustering is a dynamic clustering algorithm. Based on the principle of minimization of clustering criterion function, data is divided into different classes through iteration to make the generated classes as compact and independent as possible. In Literature [2-4], K-means algorithm and its variants are used to divide data sets in two or more ways, so as to generate decision clustering classification model. In Literature [4], the author uses the W-k-MEANS algorithm proposed in Literature [5] to cluster the data sets on non-leaf nodes, thus obtaining the decision model. The algorithm idea is as follows: firstly, let the input sample be S, and suppose the number of clusters is k, the center of each cluster C represents a class set of data clusters. Firstly, select the initial k cluster centers; then, for each sample , mark it as the cluster which closest to the center of the cluster, i.e., namely formula 4,

\[ \text{label}_i = \arg\min_{1 \leq j \leq k} \| x_i - \mu_j \| \] (4)

Finally, the center of each cluster is updated to the average value of all samples belonging to the cluster, and the new cluster center value is calculated, such as shown in formula 5.

\[ \mu_j = \frac{1}{N_j} \sum_{i=1}^{N_j} x_i \quad x_i \in C_j \] (5)

Calculate until the clustering criterion function converges, its function is as formula 6:

\[ E = \sum_{j=1}^{k} \sum_{i=1}^{N_j} \| x_i - \mu_j \|^2 \quad x_i \in C_j \] (6)

Where, E is the minimum square error of cluster division obtained by kmeans algorithm clustering, and the smaller the E value is, the higher the sample similarity in the cluster is. The time complexity of the algorithm is \( O(I*n^k) \), where I is the number of iterations and n is the data quantity. The Kmodes algorithm is a variant of the Kmeans algorithm on discrete attribute data sets, when calculating the cluster center, the mode is used instead of the mean in formula 3. When calculating the distance between the sample and the center, the distance is the number of different attributes between the sample and the cluster center. KPrototype is the combination of KMeans and Kmodes, which is used for the data set with continuous attributes and discrete attributes. When calculating the distance between the sample and the center of the cluster, the following formula 7 is adopted:

\[ \text{dist} = (1 - \alpha) \times \text{dist}_N + \alpha \times \text{dist}_C \] (7)

and represent the distance generated by the continuous attribute subset of the attribute set and the distance generated by the discrete attribute, respectively, and is the coefficient.
In this paper, in the face of continuous attribute data set, discrete attribute data set and mixed attribute data set, the clustering center generated by Kmeans, KMode and KPrototype algorithm is used to replace the category center of mass and serve as the anchor point to divide the data. The reasons for this and how to do it are explained in the next section.

2. Algorithm proposed in this paper

2.1. Basic algorithm

In this paper, the weighted classification of non-equilibrium decision tree algorithm is proposed, which divides the anchor points according to the distance from the sample to the anchor points. Firstly, the algorithm basic is given, in which the center of mass of each category sample in the training set is selected as the anchor point.

The algorithm basic:

In basic algorithm, multi-fork tree was produced. In this tree, the number of branches per non-leaf node is the same as the number of sample categories in that node. In addition to the training set D, the input of the algorithm also includes two thresholds for pre-pruning, that are minparent and mingin. These two thresholds represent the minimum number of samples of non-leaf nodes and the minimum gini index, respectively.

At the beginning of the algorithm, if the current node does not meet the conditions for division, the current node is marked as a leaf node, whose category is marked as the class with the largest sample number in the node. If the node need to be partitioned, Calculate the center of mass of each class of the current sample set as the anchor point firstly. The center of mass is calculated using the function center (ci, aj). If the property aj is a property of classification, then center gets the mode of the ci samples. Secondly, calculate the nearest anchor point for each sample xi in D, and assign the samples to the corresponding subset. The function select_near(xi,C) is to choose the nearest anchor point for sample xi. After the sample partition is complete, the function center (Di, aj) needs to be called again so as to calculate the center of each subset Di. These centers need to be kept in the tree model and used when testing invisible data.

Finally, a recursive call to the function TreeCenerate using each non-empty subset Di as input produces the next partition and returns the root node.

The basic algorithm used the nearest category center of mass as the classification criterion. Although this algorithm can generate the decision tree quickly. However, it is found that in the audio data set processed in this study, the decision tree generated by this method is usually relatively large with poor generalization ability through experiments. The reasons are as follows:

(1) The basic algorithm did not use the sample set impurity and other indicators for heuristic search, and the classification criteria were not clear;

(2) The bad effect of irrelevant and weakly correlated attributes on distance calculation is not considered.

(3) Without considering the spatial distribution of samples, sometimes the calculated category centroid cannot well represent the spatial distribution of the corresponding category. Therefore, based on the basic algorithm, there are two aspects improved in this paper, the detail description are given in 2.2 and 2.3.

2.2. Weighted for property

The classical decision tree algorithm is considered as the algorithm with attribute filtering capability. The algorithm selects the attribute most conducive to the decrease of impurity among all the attributes to divide the nodes, so these algorithms can effectively deal with irrelevant attributes and redundant attributes. However, when calculating the distance from the sample to the center of mass, all attributes are treated equally in basic algorithm. When the irrelevant attributes exist, the correlation between "the distance from the sample to the center of the category mass " and "the classification problem itself" will inevitably be weakened. Therefore, it can be concluded that if the attributes can be weighted, the
attributes related to the classification problem can be given a greater weight, otherwise a smaller weight be given. In this way, when calculating the distance, the influence of strongly correlated attributes on the distance can be amplified by weight, and the influence of weakly correlated and irrelevant attributes on the distance can be reduced, which can effectively improve the generation efficiency and accuracy of the tree.

Before the nodes are partitioned, the relief_f algorithm and the sample set in the current node are used to calculate the weight of each attribute. This is the first improvement of the algorithm proposed in this paper. Fig.1 illustrates the significance of attribute weighting with a small example. In Fig.1, the rectangle represents class 1, the circle represents category 2, the two plus signs represent the center of mass of the two categories. The solid line represents the division line generated by the unweighted attribute. The dotted line represents the dividing line generated when the X-axis attribute weight is 0.2 and the Y-axis attribute is 0.8. Obviously, the division of a category 2 sample is corrected by it.

To further illustrate the role of attribute weighting, a simple experiment was conducted. The experiment. A real-world dataset iris was used in this experiment. The 150 samples of the iris dataset were from the three categories of iris-setosa, iris-versicolor and iris-virginica, and each category has 50 samples. The data set was partitioned once using the partitioning method described in algorithmic basic. The results are as follows:

Child node 1: Iris-setosa 50
Child node 2: Iris-versicolor 44, Iris-virginica 4
Child node 3: Iris-versicolor 6, Iris-virginica 46

The number of misclassified samples is 10. And the random stratified 10% sampling is used to call the relief algorithm to obtain the weights 0.0907407, 0.138889, 0.342938 and 0.386111 for the 4 attributes. The distance obtained by weighted calculation is divided once, and the results are as follows:

Child node 1: Iris-setosa 50
Child node 2: Iris-versicolor 48, Iris-virginica 4
Child node 3: Iris-versicolor 2, Iris-virginica 46

The number of misclassified samples is 6. If partition is stopped at this point, a tree consisting of four nodes including the root node is used as the model. And the model has achieved 96% accuracy in the training set. The improvement of the basic algorithm by means of attribute weighting is called algorithm W.

2.3. The center of the cluster generated by the clustering algorithm is used as the anchor point

When samples of certain classes of the dataset are distributed in different regions of the property space, the center of a category mass is not a good representative of a sample of that category. For example, the rectangle category is distributed in two different regions in Figure 1. If the center of mass of this class and the center perpendicular line of the mass center of the circle class are used to segment the sample, the effect is obviously not good. The sample in Figure 2 clearly forms two clusters. If the sample set is divided by the central perpendicular of the Central Line of the two clusters, the samples of the five rectangular categories on the left of the graph can be separated at least.
The second improvement proposed in this paper is that on the basis of the algorithm basic, the cluster center generated by kmeans algorithm is used instead of the category of mass center as the anchor point to divide the sample. It can also be understood that the cluster generated by kmeans is the partition result of the sample set of the current node. The specific approach is to take the type of mass center as the initial cluster center of kmeans algorithm, and call this algorithm, then wait until kmeans algorithm converges or reaches the preset maximum number of iterations. In fact, this improvement is just a few repetitions of some certain steps in basic. Let's call it algorithm C. When all attributes are category attributes, kmeans algorithm transforms to kmode algorithm, the KPrototype algorithm is used when the property set contains both numeric and class properties.

2.4. The improved algorithm WDOUDT
In this paper, an integer parameter p is added to the algorithm, whose value is 0 - 3. When p is 0, the algorithm basic is used; when p contains 1, the first improvement is added; when p contains 2, the second improvement is added. In the face of specific problems to use which algorithm, determined by the user.

The above modified algorithm is called Weighted Division of Unbalanced Decision Tree, (WDOUDT for short).

Compared with basic, the algorithm WDOUDT adds parameter p to determine which improvements to apply. When p contains 1, the relief_f algorithm needs to be called to calculate the weight of each attribute. Each attribute has a weight of 1, k_means algorithm uses the centroid of each category as the initial cluster center. k_means clusters the samples into k classes. K is the number of categories in the node sample set. If the parameter p does not contain 2, Only iteration 1 round then exits. Otherwise, it needed to wait until k_means algorithm converges or reaches the maximum number of iteration rounds.

3. Experiment
3.1. Data pre-processing
This article cooperates with an organization, 800 audio data of poetry reading by school pupils were collected, five primary school teachers and broadcasters rated the frequency on a scale of five to one. Five primary school teachers and broadcasters rated the frequency on a scale of five to one. The highest score is 5 and the lowest score is 1. Five people were then rounded up for the average score of each audio as the final score for the audio. For each audio, use mel-frequency cepstral coefficients (MFCC for short) for each frame of data. But only MFCC is used as a feature, and the number of features is too small compared with the number of samples. Therefore, the minimum value, maximum value, average value, median, one quarter quantile, three quarter quantile, standard deviation and other indicators are added to the feature. After this kind of processing, the continuous audio signal is converted into the numerical data that can be processed by this algorithm.
3.2. The relationship between accuracy and time
For a single decision tree, WDOUDT, C4.5 and cart algorithms are used for comparison. In WDOUDT, the depth of decision tree is greater than 5, and the accuracy is stable; in C4.5 and cart, when the depth of decision tree is greater than 8, the accuracy is stable. The relationship between accuracy and time is shown in Table 1.

Table 1. The accuracy and time length of the three algorithms are compared.

| Algorithm | Accuracy (%) | Depth of Tree | Time Length (hours) |
|-----------|--------------|---------------|---------------------|
| WDOUDT    | 78.1         | 5             | 2.3                 |
| C4.5      | 75.3         | 8             | 4.1                 |
| CART      | 76.1         | 8             | 3.8                 |

3.3. The cost of generating random forest
As same as other decision tree algorithms, the WDOUDT algorithm proposed in this paper is not very accurate for the converted audio data, which is far from the product standard. Then bagging and boosting are usually used to improve the accuracy, random forest is one of the most commonly used methods. Next, three different algorithms are used to construct decision tree and generate random forest. The relationship between accuracy and the number of decision trees is shown in Figure 3.

![Figure 3. Relationship between accuracy and number of decision trees](image)

![Figure 4. Comparison of the time required to achieve the same accuracy](image)

The experimental environment is as follows: CPU Intel Core i7-6500U, 2 core, 2.5GHz per core, RAM 16GB. test the time required for different algorithms to achieve the same accuracy, which as shown in Figure 4. From the above results, WDOUDT algorithm can get better results than C4.5 and CART algorithm, and it takes less time to generate a single decision tree or random forest. However, the accuracy of the three algorithms is not very high. The reason for this is that the analysis of classification error data of the three algorithms mainly focuses on the level 1 and level 5 classification. Due to the low scores of 1 and 5 in the original data (especially 1 point data, less than 10 items), their common characteristics are not so uniform and obvious. And the accuracy of 3 and 4 scores is almost 100%, because the original data is the most abundant and the samples are the most abundant.

4. Summary and Prospect
In this paper, a kind of weighted partition non-equilibrium decision tree (WDOUDT) was proposed,
this algorithm can be effectively applied to the field of poetry reading emotion recognition. The WDOUDT proposes two improvements based on decision tree: the first is Attribute weighting, and the second is using the anchor point instead of the center of mass as the cluster center. Experiments show that, WDOUDT has better precision than C4.5, CART and other classic decision trees. WDOUDT is suitable for large amount of data and multiple samples.

Acknowledgments
First and foremost, I would like to show my deepest gratitude to Professor Zhang Dongqing, a respectable, and resourceful scholar, who has provided us with valuable guidance in every stage of the writing of this thesis, we also acknowledge financial support from the Dr. startup Fund of Liaoning Province (20170520398) and Science and technology general project of Liaoning Provincial Department of Education (l2015041). Last but not least, I’d like to thank all my friends and my colleagues for their encouragement and support.

References
[1] Kononenko I. Estimating attributes: Analysis and extensions of RELIEF[M]// Machine Learning: ECML-94. Springer Berlin Heidelberg, 1994.
[2] Mui J K, Fu K S. Automated classification of nucleated blood cells using a binary tree classifier[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1980, PAMI-2(5):429-443.
[3] Huang Z, Ng M, Lin T, et al. An interactive approach to building classification models by clustering and cluster validation[J]. 2000.
[4] Li Y, Hung E, Chung K, et al. USING A VARIABLE WEIGHTING k-MEANS METHOD TO BUILD A DECISION CLUSTER CLASSIFICATION MODEL[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2012, 26(02):1250003.
[5] Huang J, Ng M, Rong H, et al. Automated Variable Weighting in k-Means Type Clustering[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2005, 27(5):657-668.