An Improved Approach to Ordinal Classification

Donghui Wang, Junhai Zhai(✉), Hong Zhu, and Xizhao Wang

Key Laboratory of Machine Learning and Computational Intelligence,
College of Mathematics and Computer Science, Hebei University,
Baoding 071002, China
mczjh@126.com

Abstract. A simple ordinal classification approach (SOCA) has been proposed by Frank and Hall. SOCA is a general method, any classification algorithm such as C4.5, k nearest neighbors (KNN) algorithm and extreme learning machine (ELM) etc. can be applied to this approach. We find that in SOCA only ordering information of decision attribute is used to classify objects but the ordering information of conditional attributes is not considered. Furthermore we experimentally find that ordering information of conditional attributes can also improve the generalization ability of the classification method. In this paper, we propose an improved ordinal classification methodology by employing the ordering information of both condition and decision attributes. In addition, we analyze the sensitivity of the SOCA on performance to the underlying classification algorithms, for instance, C4.5, KNN and ELM. A number of experiments are conducted and the experimental results show that the proposed method is feasible and effective.

Keywords: Ordinal classification · Monotonic classification · Decision tree · Rank mutual information

1 Introduction

General classification algorithms do not consider the ordering information including the conditional attributes and the decision attribute. Actually, the ordering information can make contribution to classification. Classification problems with ordering information are called ordinal classification problems, which are also named monotonic classification problems.

As early as 1989, David et al. [1] have studied the ordinal classification problems, and proposed an approach for learning and classification of monotonic ordinal concepts. In the next few years, David developed other two algorithms for monotonic classification problems [2,3]. In literature [2], ordinal classification for multi-attribute decision making has been studied for discrete domains. In reference [3], an information-theoretic machine learning algorithm with monotonicity maintenance was proposed.

Since the pioneering work of David, the ordinal classification problems have been investigated by many machine learning researchers. Potharst and Bioch [4] proposed an order preserving tree-generation algorithm for multi-attribute classification problems with k linearly ordered classes and an algorithm for repairing non-monotonic decision trees. Frank and Hall [5] provide a general method

X. Wang et al. (Eds.): ICMLC 2014, CCIS 481, pp. 33–42, 2014.
DOI: 10.1007/978-3-662-45652-1_4

SOCA for classification of data sets with ordering information in decision attribute, which partition the k-class ordinal problem into (k-1)-binary class problems. SOCA enables standard classification algorithms, such as, C4.5, KNN, and ELM etc. to exploit the ordering information. The key superiority of SOCA is that it does not require any modification of the underlying learning algorithms. However, SOCA ignores the ordering information in conditional attributes. Feelders and Pardoel [6] studied the pruning problem of monotone classification trees, and proposed a pruning algorithm which can improve the calculation efficiency of induction of trees. Cardoso and Ricardo [7] investigated the criteria for measuring the performance of ordinal classification, and based on confusion matrix, proposed a new metric. Wojciech and Slowinski [8] discussed the problem of nonparametric ordinal classification with monotonicity constraints and presented a statistical framework for classification with monotonicity constraints. The main contribution of [8] is that a statistical theory for ordinal classification with monotonicity constraints is developed. Hu et al. [9,10] extended the ideas of ID3 algorithm to the monotonic classification, and investigated the corresponding heuristic, i.e., rank mutual information, and the corresponding tree generation algorithm.

Recently, a number of new problems related to ordinal classification have attracted the attention of many machine learning researchers. One example is that in [11,12] Hu et al studied the feature selection problem for monotonic classification. Based on their proposed rank mutual information, a feature selection algorithm was developed. Another example is that, based on cluster and variability analyses, Lin [13] proposed a feature selection algorithm for ordinal multi-class classification problems.

In addition, the cost-sensitive based ordinal classification problems were studied in [14–16]. Moewes and Kruse [18] extended the ordinal classification to fuzzy scenario, and proposed a fuzzy ordinal classification method. Many applications of ordinal decision trees can be found from references. For example, Baccianella et al. in [17] applied the ordinal classification technique to text information mining and obtained a promising performance.

From the literature we do not find a study on the extension of SOCA by considering the ordering information of conditional attributes, aiming at an improvement of accuracy for ordinal classification. In this paper, we make an attempt to conduct such an improvement. By considering the ordering information both in conditional attributes and decision attribute, we develop an improved model of ordinal classification SOCA. We analyze the sensitivity of the SOCA on performance to the underlying classification algorithms, for instance, C4.5, KNN and ELM. A number of expe-riments are conducted and the experimental results show that the proposed method is feasible and effective. The experimental results also show that our proposed approach is not sensitive to the underlying algorithms.

The paper is organized as follows. Section 2 provides some necessary related preliminaries. Section 3 presents our improved model of SOCA by considering the ordering information of both conditional attributes and decision attribute. The experimental results and their corresponding analysis are listed in Section 4. Section 5 concludes this paper.

2 Preliminaries

2.1 Transform $k-$Class Ordinal Classification Problem to $(k-1)-$Binary Classification Problem

SOCA is a kind of classification methodologies which transform a $k-$ class ordinal classification problem into $(k-1)-$binary classification problems. The main process is to convert the class attribute C^* with ordered values $C_1, C_2, \cdots, C_k$ into $k-1$ binary attributes. The original dataset can be transformed into k-1 derived dataset. The derived datasets contain the same attribute values for each instance, and the derived binary class value is transformed from original class value by using the rule:

If $C(x) > C_i$ then $C - binary_i(x) = 1$; else $C - binary_i(x) = 0$. where x is a sample, $C(x)$ is the value of original class, and $C - binary_i(x)$ is the value of ith derived binary class.

Based on each derived dataset, we can get a binary classifier, then we can get k-1 binary classifiers.

In the process of testing, each sample is processed by each of the $k-1$ binary classifiers and then the probabilities $P_1, P_2, \cdots, P_{k-1}$ are obtained,where P_i represents the decision value of sample x in ith binary classifier.

The probability of each of the k ordinal decision values is calculated using the following formulas.

$$P(C_1) = 1 - P_1 \tag{1}$$

$$P(C_i) = P_{i-1} - P_i, 1 < i < k \tag{2}$$

$$P(C_k) = P_{k-1} \tag{3}$$

The class with maximum probability is assigned to the sample.

2.2 Ranking Information Entropy and Ranking Mutual Information

For the ordinal classification problems with monotonicity constraints, based on rank mutual information, an ordinal decision tree algorithm (REMT) was proposed [9]. In comparison with the general ordinal classification model which assumes that condi-tion attributes and decision attribute are all ordinal or only the decision is ordinal, the current problem assumes an additional monotonicity constraint, given two samples x and y, if $x \leq y$, then we have $f(x) \leq f(y)$.

Definition 1. *Let $U = \{x_1, x_2, \cdots, x_n\}$ be a set of objects, $B \subseteq A$ where A is a set of attributes. The upwards ranking entropy is defined as*

$$RH_B^{\geq}(U) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{|[x_i]_B^{\geq}|}{n} \tag{4}$$

Similarity the downwards ranking entropy of the set U is defined as

$$RH_B^{\leq}(U) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{|[x_i]_B^{\leq}|}{n} \tag{5}$$

Definition 2. *Let U be a set of objects described with a set of attributes A, $B \subseteq A$, $C \subseteq A$. The upwards ranking mutual information of B and C is defined as*

$$RMI^{\geq}(B,C) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{|[x_i]_B^{\geq}| \times |[x_i]_C^{\geq}|}{n \times |[x_i]_B^{\geq} \cap [x_i]_C^{\geq}|} \tag{6}$$

and downwards ranking mutual information of B and C is defined as

$$RMI^{\leq}(B,C) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{|[x_i]_B^{\leq}| \times |[x_i]_C^{\leq}|}{n \times |[x_i]_B^{\leq} \cap [x_i]_C^{\leq}|} \tag{7}$$

3 The Improved Approach

In this section, we first conduct an analysis on the sensitivity of SOCA about its performance to the underlying classification algorithms, and then present the improved approach.

3.1 The Analysis of SOCA Sensitivity

Frank and Hall claim that SOCA is applicable in conjunction with any base learner that can output class probability estimates. The performance of SOCA may change with the change of underlying classifier, in this paper, we analysis the SOCA sensitivity by applying C4.5, KNN and ELM to SOCA respectively.

In C4.5, the value of leaf node is defined as the probability of samples whose decision value is 1 in current leaf node. Then, to a testing sample, we can get its class probability estimates.

In KNN, the decision value is defined as the probability of samples whose decision value is 1 in the k nearest neighbor. Due to KNN can output class probability estimates, we can apply SOCA to it.

In ELM, the decision value of a sample is defined as the degree of the sample belong to class 1. Then we transform the degree to a probability .In this way, ELM can be applied to SOCA.

3.2 Our Improved Approach

According to SOCA, we transform $k-$class ordinal classification problem with monotonicity constraint to $(k-1)$binary-class ordinal classification problems. Then we use REMT to solve the $(k-1)$binary-class ordinal classification problems. In other words, we can get $k-1$ derived datasets based on SOCA, and then execute the process of building tree with REMT to each derived dataset. The details of the process are as follows.

1. In the process of building tree, in order to create nodes for the tree, we need to select $a_{\max}$ and $c_{\max}$ that satisfy $MaxRMI_{c_j} = RMI(a_i, D)$ by computing the maximal ranking mutual information between candidate attributes and decision attribute.a_i is the candidate attribute, c_j is attribute values of candidate a_i. Then

we set $a_{\max}$ as the extended node and divide the node base on $c_{\max}$. However, if ranking mutual information of node attributes less than a specific value, or all of the samples are classified to the same class, we need to generate a leaf node base on the current samples. Because that the decision value is 0 or 1. The probability of the samples whose decision values are 1 is assigned to the current leaf node. For example: current leaf node contains 5 training samples, the decision value of 3 samples is 1, and the others are 0. Then we assign 3/5 as the leaf node's decision value. Thus, k ordinal classification problem with monotonicity constraint can attains $k-1$ binary trees through the above method.

2. In the process of searching tree, each sample is processed by each of the $k-1$ binary tree and obtain the probabilities $P_1, P_2, \cdots P_{k-1}$(the decision value of current sample in $k-1$ binary tree). The probability assigned to the k decision values $(C_1, C_2, \cdots, C_k)$ of the sample is calculated using the following formulas.

$$P(C_1) = 1 - P_1 \tag{8}$$

$$P(C_i) = P_{i-1} - P_i, 1 < i < k \tag{9}$$

$$P(C_k) = P_{k-1} \tag{10}$$

The class with the maximum probability is assigned to the sample.

4 Experimental Results

In order to verify the effectiveness of the proposed approach, we conduct some experiments with 10-fold cross validation on UCI datasets and artificial monotonicity constraint datasets. The basic information of the selected UCI datasets is listed in Table 1. The experimental environment is PC with 2.2GHz CPU and 8G memory, the operating system is Windows 7, MATLAB 7.1 is the experimental platform.

Table 1. The basic of the selected UCI datasets

datasets	#attributes	#classes	#samples
Abalone	8	29	4177
Cancer	10	2	341
Car	6	4	864
Mushroom	22	2	2820
Diabetes	8	2	576
German	20	2	500
Ionosphere	34	2	176
Sat	36	7	4290
Segment	19	7	1540
Servo	4	44	111
Sonar	60	2	150
Wave	21	3	3332

4.1 Sensitivity Analysis of SOCA

In order to analyze the sensitivity, we apply SOCA to C4.5, KNN and ELM respectively on UCI datasets. The average performance is depicted in Figure 1.

We use paired two sided t-test to test that if there is significance between the two results obtained with C4.5, KNN, ELM and with SOCA+C4.5, SOCA+KNN, SOCA+ELM respectively. The significance level is set to 1%. The results are given in Table 2. Through the experiments, we found the SOCA is not sensitive to the underlying classifier.

Table 2. The results of paired two sided t-test

results	C4.5	KNN	ELM
improvement	7(4)	4(1)	10(5)
degradation	2(0)	2(0)	2(2)

4.2 Performance of Our Proposed Approach

We conduct another experiment on an artificial data set to analyse the performance of the poposed approach. The artificial data set are generated with the following formula:

$$f(x_1, x_2) = x_1 + \frac{1}{2}(x_2^2 - x_1^2) \tag{11}$$

Where, x_1 and x_2 are two random variables which are independently drawn from [0, 1]. In order to generate ordered class labels, the resulting values were discretized into k intervals $[0, 1/k], (1/k, 2/k], \cdots, (k-1/k, 1]$. Thus each interval contains approximately the same number of samples. The samples belonging to one of the intervals share the same decision value. Then we form a $k-$class monotonic classification task. In this experiment, we try k =2, 4, 6, 8, 10, 20 and 30 respectively.

We first study the proposed approach on different numbers of classes. We generate a set of artificial datasets with 1000 samples and the numbers of classes vary from 2 to 30. Based on 5-fold cross validation technique, the experiment was repeated 100 times. The average performance is computed and given in Table 3; the curves of testing accuracy with number of classes are shown in Figure 2. The proposed method yields the better testing accuracy in all the cases except the case two classes are considered.

In addition, we also consider the influence of sample numbers on the performance of trained models. For artificial data of 1000 samples with 6, 10, 20, 30 classes respectively, we randomly draw training samples from a dataset. The size of training samples ranges from 50 to 200. In this process, we guarantee there is at least one representative sample from each class.

The rest samples are used in testing for estimating the performance of the trained models. To the different classes, the curves of testing accuracy with

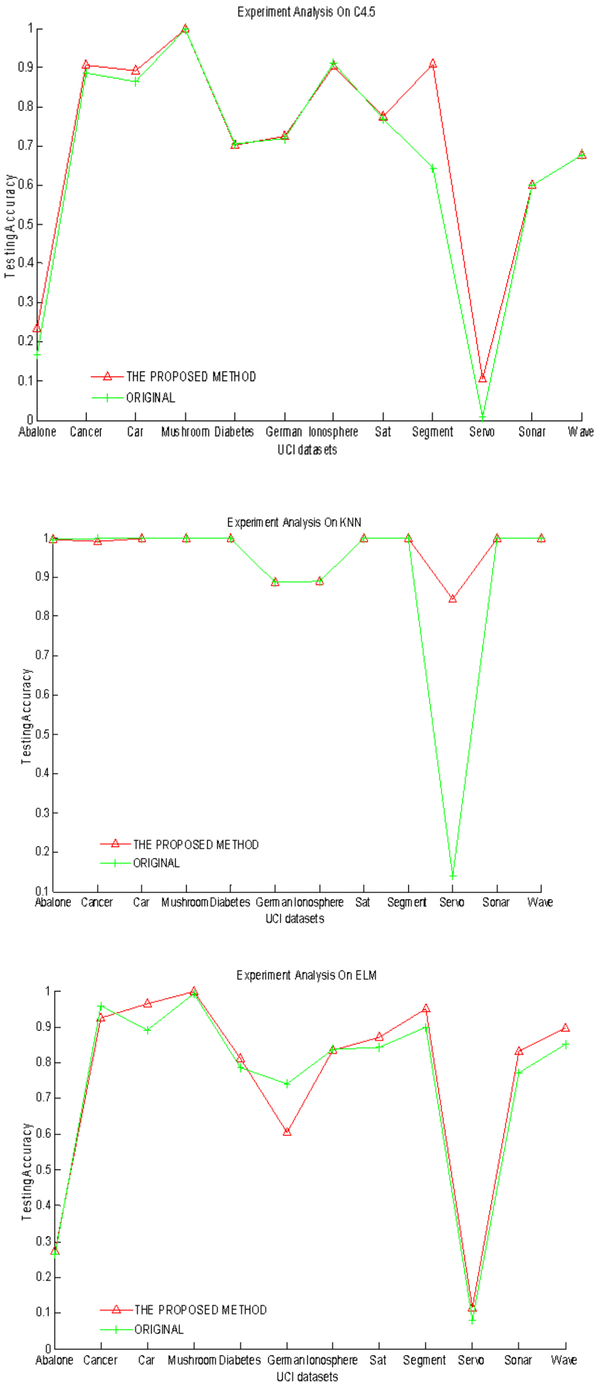

Fig. 1. The experimental results with different methods: C4.5, KNN and ELM

Table 3. The testing accuracy on artificial data with the number of class values vary from 2 to 30

Classes	REMT	The proposed method
2	0.9624	0.9626
4	0.9126	0.9096
6	0.9113	0.9043
8	0.8507	0.8210
10	0.8010	0.7984
20	0.7173	0.6859
30	0.5375	0.5054

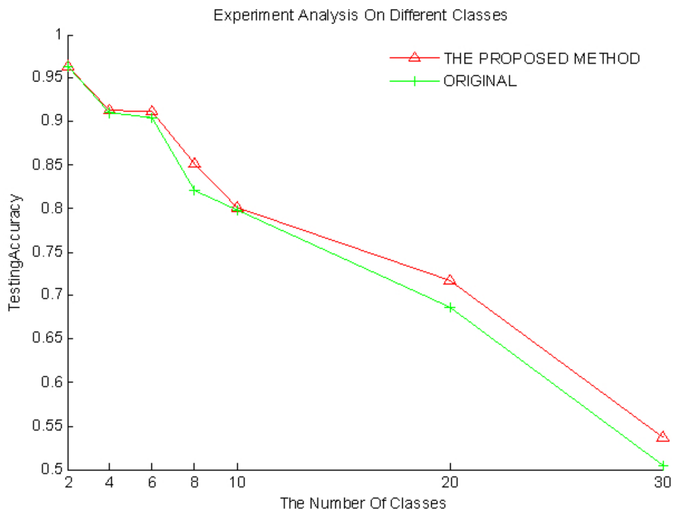

Fig. 2. The testing accuracy on artificial data with different classes

number of training samples change are shown in Figure 2. We can see that the accuracy of the proposed method is higher than REMT, no matter how many training samples are used.

5 Conclusions

We extended the method SOCA, and proposed an improved ordinal classification approach in this paper. The ordering information both of conditional attributes and decision attribute are all be taken into consideration in the improved approach. We also analysed the sensitivity of the SOCA on performance to the underlying classification algorithms, and obtained the conclusion that SOCA is not sensitive to the underlying algorithms. In our future works, we will study whether any ordinal classifica-tion algorithm whose outputs are posteriori probabilities can be used as the underlying algorithms, not only limited to the decision tree algorithm.

Acknowledgments. This research is supported by the national natural science foundation of China (61170040, 71371063), by the key scientific research foundation of education department of Hebei Province (ZD20131028), by the scientific research foundation of education department of Hebei Province (Z2012101), and by the natural science foundation of Hebei Province (F2013201110, F2013201220).

References

1. Ben-David, A., Sterling, L., Pao, Y.H.: Learning and classification of monotonic ordinal concepts. Comp. Intell. **5**, 45–49 (1989)
2. Ben-David, A.: Automatic generation of symbolic multiattribute ordinal knowledge-based DSSs: methodology and applications. Decision Sciences **23**, 1357–1372 (1992)
3. Ben-David, A.: Monotonicity maintenance in information-theoretic machine learning algorithms. Machine Learning **19**, 29–43 (1995)
4. Potharst, R., Bioch, J.C.: Decision trees for ordinal classification. Intelligent Data Analysis **4**, 99–111 (2000)
5. Frank, E., Hall, M.: A simple approach to ordinal classification. In: Flach, P.A., De Raedt, L. (eds.) ECML 2001. LNCS (LNAI), vol. 2167, pp. 145–156. Springer, Heidelberg (2001)
6. Feelders, A., Pardoel, M.: Pruning for monotone classification trees. In: Berthold, M., Lenz, H.-J., Bradley, E., Kruse, R., Borgelt, C. (eds.) IDA 2003. LNCS, vol. 2810, pp. 1–12. Springer, Heidelberg (2003)
7. Cardoso, J.S., Ricardo, S.: Measuring the performance of ordinal classification. International Journal of Pattern Recognition and Artificial Intelligence **25**, 1173–1195 (2011)
8. Wojciech, K., Slowinski, R.: On nonparametric ordinal classification with monotonicity constraints. IEEE Transactions on Knowledge and Data Engineering **25**, 2576–2589 (2013)
9. Hu, Q., Che, X., Zhang, L.: Rank entropy-based decision trees for monotonic classification. IEEE Transactions on Knowledge and Data Engineering **24**, 2052–2064 (2012)
10. Hu, Q., Guo, M., Da, R.: Information entropy for ordinal classification. Science China-Information Sciences **53**, 1188–1200 (2010)
11. Hu, Q., Pan, W., Song, Y.: Large-margin feature selection for monotonic classification. Knowledge-Based Systems **31**, 8–18 (2012)
12. Hu, Q., Pan, W., Zhang, L.: Feature Selection for Monotonic Classification. IEEE Transactions on Fuzzy Systems **20**, 69–81 (2012)
13. Lin, H.Y.: Feature selection based on cluster and variability analyses for ordinal multi-class classification problems. Knowledge-Based Systems **37**, 94–104 (2013)
14. Ruan, Y.X., Lin, H.T., Tsai, M.F.: Improving ranking performance with cost-sensitive ordinal classification via regression. Information Retrieval **17**, 1–20 (2014)
15. Lin, H.T., Li, L.: Reduction from cost-sensitive ordinal ranking to weighted binary classification. Neural Computation **24**, 1329–1367 (2012)

16. Kotsiantis, S.B., Pintelas, P.E.: A cost sensitive technique for ordinal classification problems. In: Vouros, G.A., Panayiotopoulos, T. (eds.) SETN 2004. LNCS (LNAI), vol. 3025, pp. 220–229. Springer, Heidelberg (2004)
17. Baccianella, S., Esuli, A., Sebastiani, F.: Using micro-documents for feature selection: The case of ordinal text classification. Expert Systems with Applications **40**, 4687–4696 (2013)
18. Moewes, C., Kruse, R.: Evolutionary fuzzy rules for ordinal binary classification with monotonicity constraints. Soft Computing **291**, 105–112 (2013)