A Lazy One-Dependence Classification Algorithm Based on Selective Patterns

Zhuoya Ju[1], Zhihai Wang[1(✉)], and Shiqiang Wang[2]

[1] Beijing Jiaotong University, Beijing 100044, China
{juzhuoya,zhhwang}@bjtu.edu.cn
[2] 1Verge Internet Technology (Beijing) Co., Ltd., Beijing, China
wangshiqiang@alibaba-inc.com

Abstract. Data mining is a widely acceptable method on mining knowledge from large databases, and classification is an important technique in this research field. A naïve Bayesian classifier is a simple but effective probabilistic classifier, which has been widely used in classification. It is commonly thought to assume that the probability of each attribute belonging to a given class value is independent of all other attributes in the naïve Bayesian classifier; however, there are lots of contexts where the dependencies between attributes are complex and should thus be considered carefully. It is an important technique that constructing a classifier using specific patterns based on "attribute-value" pairs in lots of researchers' work, and the classification result will be impacted by dependencies between these specific patterns meanwhile. In this paper, a lazy one-dependence classification algorithm based on selective patterns is proposed, which utilizes both the patterns' discrimination and dependencies between attributes. The classification accuracy benefits from mining and employing patterns which own high discrimination, and building the one-dependence relationship between attributes in a proper way. Through an exhaustive experimental evaluation, it shows that the proposed algorithm is competitive in accuracy with the state-of-the-art classification techniques on datasets from the UCI repository.

Keywords: Classification · Pattern discovery · Dependence Bayesian classifier · Lazy learning

1 Introduction

In the machine learning field and data mining technology, classification is regarded as a crucial learning method. Classification algorithms based on Bayesian network own a solid theoretical basis, strong anti-noise performance, good classification performance and robustness. A naïve Bayesian classifier is commonly thought to assume that the probability of each attribute belonging

Supported by Beijing Natural Science Foundation (4182052), and National Natural Science Foundation of China (61672086, 61771058).

X. Geng and B.-H. Kang (Eds.): PRICAI 2018, LNAI 11013, pp. 113–120, 2018.
https://doi.org/10.1007/978-3-319-97310-4_13

to a given class value is independent of all other attributes, and this assumption is so called "conditional independence assumption" [1]. However, there are lots of contexts where the dependencies between attributes are complex and should thus be considered carefully. To accord with actual situation, dependencies between attributes are researched deeply in many classification algorithms: Tree Augmented Naïve Bayes Classifier (TAN) [2], Aggregating One-Dependence Estimators (AODE) [3], and its extended algorithms [4–6], and so on. Nevertheless, it does not delve into that the key role of patterns based on attributes and their values in these classification algorithms.

In general, an instance is characterized by n ($n = 1$, 2, 3, ...) pairs of "attribute-value" which are called "items". Each instance without missing values can be considered as an itemset with n items. There has been a great deal of research on the use of itemsets to complete the classification tasks and achieved high classification accuracy already. For instance, Classification by Aggregating Emerging Pattern (CAEP) [7], the Bayesian Classification based on Emerging Patterns (BCEP) [8], classifiers based on Jumping Emerging Patterns (JEPs) [9], and so on.

In order to exploit the patterns' discrimination and construct dependencies between attributes using Bayesian networks, this paper proposes a lazy one-dependence classification algorithm based on the selective patterns. The selective patterns (such as frequent patterns, emerging patterns, etc.) that play major roles as the basis for classification are mined first, and then two types of attributes (belonging to the selective patterns or not) are analyzed by the Bayesian network which constructs a one-dependence relationship. After that a new classification model is built.

2 Background

The classification based on the selective patterns is to find out some specific patterns with high growth rates from non-target to target class, and analyze dependencies between the attributes contained in these specific patterns and other attributes. Some rudimentary definitions and formulas are given below.

The *support* of itemset *I*, $Supp_D(I) = count_D(I)/|D|$, where $count_D(I)$ is the number of instances in dataset D that contains itemset *I*, and $|D|$ is the total number of instances in dataset D.

Given a dataset D, which is divided into two different subsets D_1 and D_2, namely, $D_1 \cap D_2 = \emptyset$, $D_1 \cup D_2 = D$. The *growth rate* of itemset I from D_1 to D_2, namely $Growth(I, D_1, D_2)$, is defined as follows,

$$Growth(I, D_1, D_2) = \begin{cases} 0 & Supp_{D_1}(I) = Supp_{D_2}(I) = 0 \\ \infty & Supp_{D_1}(I) = 0, Supp_{D_2}(I) > 0 \\ \frac{Supp_{D_2}(I)}{Supp_{D_1}(I)} & others \end{cases} \quad (1)$$

A *Selective pattern* is the itemset I whose support $Supp_{D_1}(I)$ and growth rate $Growth(I, D_1, D_2)$ satisfy the threshold ξ, ρ respectively, and thus it owns discrimination to classify instances.

Patterns represent the nature and important characteristics of datasets and form the basis of many important data mining tasks. The boundary algorithm BORDER_DIFF proposed by Dong et al. [10] is used to mine specific patterns for classification in this paper. Removal of redundant patterns and noises will help to speed up the classification and improve the classification accuracy. In this paper, the method in [8] is used to filtering patterns.

The aggregation one-dependence estimator (AODE) averages all models from a restricted class of one-dependence classifiers, the class of all such classifiers that have all other attributes depend on a common attribute and the class [3]. Each instance can be described by an n-dimensional attribute vector $X = (a_1, a_2, \ldots, a_n)$, where a_i represents the value of the ith attribute A_i. Given instance X, the task of classification is to calculate the class of the maximum a posteriori (MAP) as X's prediction class, which can thus be expressed as:

$$P(c_k|X) \propto \frac{\sum_{i:1\leq i\leq n\wedge F(a_i)\geq u} P(c_k, a_i)\cdot \prod_{j=1,j\neq i}^{n} P(a_j|c_k, a_i)}{|i:1\leq i\leq n\wedge F(a_i)\geq u|} \tag{2}$$

where $F(a_i)$ is a count of the number of training examples having attribute-value a_i and is used to enforce the limit u placed on the support needed in order to accept a conditional probability estimate.

3 A Lazy Classification Algorithm Based on Selective Patterns

According to Bayes theorem, $P(c|X)$ can be expressed as:

$$P(c|a_1, a_2, \ldots, a_n) = \frac{P(a_1, a_2, \ldots, a_n|c)P(c)}{P(a_1, a_2, \ldots, a_n)} \propto P(a_1, a_2, \ldots, a_n|c)P(c) \tag{3}$$

Given a class c, assuming that the attributes contained in the pattern and other attributes are conditionally independent of each other, namely:

$$\begin{aligned} P(c|a_1, a_2, \ldots, a_n) &\propto P(c)P(a_1, a_2, \ldots, a_i|c)P(a_{i+1}, \ldots, a_n|c) \\ &= P(c|a_1, a_2, \ldots, a_i)P(a_1, a_2, \ldots, a_i)P(a_{i+1}, \ldots, a_n|c) \end{aligned} \tag{4}$$

where $\{a_1, a_2, \ldots, a_i\}$ represents the attributes included in the pattern.

3.1 Characterization of Discriminative Patterns

Assume that the set of attributes contained in itemset e is $\{a_1, a_2, \ldots, a_i\}$, which is a special pattern of a growth rate of $Growth(e, c', c)$ from a data set of class c' to a data set of class c. When a test instance contains e, the probability that this instance belongs to class c is $P(c|a_1, a_2, \ldots, a_i)$, and abbreviated as $P(c|e)$. According to the definitions of *support* and *growth rate*, then:

$$P(c|e) = \frac{Growth(e, c', c)}{Growth(e, c', c)\frac{|c|}{|c'|} + 1} \frac{|c| + |c'|}{|c'|} P(c) \tag{5}$$

3.2 A Lazy One-Dependence Classification Algorithm

The Aggregate One-Dependence Classification based on Selective Patterns (AODSP) is proposed in this paper. Attributes in a specific pattern are treated as a whole, and the attributes in the pattern are assumed to be independent of attributes out of the pattern. AODSP assumes that the dependencies between attributes out of a specific pattern satisfy a one-level Bayesian tree structure, that is, each attribute sequentially serves as the parent of other attributes and the remaining attributes depend on this attribute as child nodes. The average probability calculated by the multiple classifiers is as the classification probability.

Let E be a set of all patterns, and the attributes in a pattern e which is contained in E are treated as a whole. From the Eq. 2, the conditional probability of attributes out of e satisfies:

$$\begin{aligned} P(a_{i+1},\ldots,a_n|c) &= \frac{P(c|a_{i+1},\ldots,a_n)P(a_{i+1},\ldots,a_n)}{P(c)} \propto \frac{P(c|a_{i+1},\ldots,a_n)}{P(c)} \\ &\propto \frac{\sum_{j:i+1\leq j\leq n\wedge F(a_j)\geq u} P(c,a_j)\prod_{k=i+1,k\neq j}^{n} P(a_k|c,a_j)}{|j: i+1\leq j\leq n\wedge F(a_j)\geq u|}\frac{1}{P(c)} \end{aligned} \tag{6}$$

and then:

$$\begin{aligned} P(c|a_1,a_2,\ldots,a_n) &\propto P(c|a_1,a_2,\ldots,a_i)P(a_1,a_2,\ldots,a_i)P(a_{i+1},\ldots,a_n|c) \\ &\propto P(c|a_1,a_2,\ldots,a_i)P(a_1,a_2,\ldots,a_i) \\ &\cdot \frac{\sum_{j:i+1\leq j\leq n\wedge F(a_j)\geq u} P(c,a_j)\prod_{k=i+1,k\neq j}^{n} P(a_k|c,a_j)}{|j: i+1\leq j\leq n\wedge F(a_j)\geq u|}\frac{1}{P(c)} \end{aligned} \tag{7}$$

The patterns' discrimination is applied to the Bayesian network, that is, the Eq. 5 is substituted into the Eq. 7 and the discrimination of all patterns is aggregated to obtain the probability prediction equation adopted by the aggregation one-dependent classification algorithm based on selective patterns:

$$\begin{aligned} P(c|a_1,a_2,\ldots,a_n) \propto \sum_{e\in E}\Big(&\frac{Growth(e,c',c)}{Growth(e,c',c)\frac{|c|}{|c'|}+1}\frac{|c|+|c'|}{|c'|}\cdot P(a_1,a_2,\ldots,a_i) \\ &\cdot\frac{\sum_{j:i+1\leq j\leq n\wedge F(a_j)\geq u} P(c,a_j)\prod_{k=i+1,k\neq j}^{n} P(a_k|c,a_j)}{|j: i+1\leq j\leq n\wedge F(a_j)\geq u|}\Big) \end{aligned} \tag{8}$$

Equation 8 aggregates the discrimination of all patterns, and further considers the dependencies between attributes out of patterns. When classifying a test instance, the class of the instance will be the class that maximizes the Eq. 8.

The AODSP algorithm is defined in Algorithm 1.

4 Experiments and Evaluations

In order to validate the accuracy of the proposed aggregation one-dependent classification algorithm based on the selective patterns, 8 datasets from the UCI repository of machine learning databases [11] are used as experimental datasets (see Table 1).

Algorithm 1. AODSP($instance, m, E$)

```
Input:
    instance: to be classified; m: the number of class labels; E: sets of patterns.
Output:
    probs: distribution of prediction class labels of instance.
for $i \in [0, m]$ do
    $probs[i] = 0$;
    for $j \in [0, E[i].size()]$ do
        Itemset $e$ = (Itemset)$E[i]$.elementAt($j$);
        if $instance$ contains $e$ then
            calculate $P(i|\text{attributes in } e)$;
            calculate $P_{AODE}(\text{attributes not in } e|i)$;
            calculate $P(a_1, a_2, \ldots, a_i)$, where $a_1, a_2, \ldots, a_i \in e$;
            $p = P(i|instance)$;
        end if
        $probs[i] = probs[i] + p$;
    end for
    if $probs[i] \leq 0$ then
        $probs[i]$=aode.distributionForInstance($instance$);
    end if
end for
Normalize $probs$;
return $probs$;
```

4.1 Parameter Analysis

Patterns' discrimination is determined by the growth rate and the supports in different classes. Table 2 shows the error rates of AODSP on the iris dataset with different supports. Wherein, the value of "S" means the random seed used in the experiment; "Mean" means the average value of 5 results; "Support1" and "Support2" respectively represent the support of patterns on the non-target class and the target class; "Growth Rate" means the growth rate from non-target class to target class.

It can be figured out that: when the support is set too high, selective patterns can't be found, AODSP degenerates to AODE; when the support is set too low, there will be too many patterns which may lead to overfitting and the error rate is compromised.

4.2 Empirical Setup

For numerical attributes, the multi-interval discretization method provided in [12] is adopted to discrete them as a preprocessing step. The experiment uses a 10-fold cross-validation method to calculate the error rate of the classifiers. The random seed is set as 1, 2, 3, 5, and 7 respectively to calculate the error rate, and the average is taken as classification results. The AODSP is compared with the NB, AODE, ASAODE, and BCEP in error rate. AODSP takes the

Table 1. Summary of datasets.

No	Domain	Data file	Instances	Attributes	Classes	Missing value
1	Balance Scale	balance-scale	625	4	3	No
2	Liver Disorders	bupa	345	6	2	No
3	German	german	1000	20	2	No
4	House Votes 84	house-votes-84	435	16	2	No
5	Iris Classification	iris	150	4	3	No
6	Labor Negotiations	labor	57	16	2	Yes
7	New-Thyroid	new-thyroid	215	5	3	No
8	Pima Indians Diabetes	pid	768	8	2	No

Note: The number of "Attributes" does not include the class attribute.

Table 2. Error rates of AODSP on iris with respect to different supports.

No	$S = 1$	$S = 2$	$S = 3$	$S = 5$	$S = 7$	Mean	Support1	Support2	Growth rate
1	6.67	6.67	7.33	6.00	6.00	6.53	0.05	0.8	16
2	5.33	6.00	5.33	4.00	6.00	5.33	0.05	0.5	10
3	4.67	4.67	4.67	4.67	4.67	4.67	0.1	0.5	5
4	5.33	5.33	5.33	6.00	5.33	5.47	0.2	0.4	2
5	6.67	6.67	7.33	7.33	6.00	6.80	0.4	0.6	1.5
6	6.67	6.67	7.33	7.33	6.00	6.80	0.5	0.6	1.2

support and growth threshold the same as BCEP (namely, the minimum support threshold $\xi = 1\%$ or an absolute count of 5; the minimum growth rate $\rho = 5$).

4.3 Error Rate Analysis

Table 3 shows the error rates of the 5 classifiers on 8 data sets, with the lowest error rate in bold. The AODSP based on selective patterns proposed in this paper adopts the idea of AODE to deal with the dependencies between attributes in and out of specific patterns. ASAODE is an improvement of AODE, and BCEP tries to combine emerging patterns with Bayesian network. They all belong to Bayesian network and are picked as reference objects.

According to the experiment, the following conclusions can be figured out: Compared to AODE, AODSP achieves lower error rate on 5 datasets and the same error rate on 1 datasets; Compared to BCEP, AODSP achieves lower error rate on all 8 datasets; Compared to other classifiers, AODSP achieves the lowest average error rate on all 8 datasets.

The experiment uses the method proposed by Demšar [13] to test the critical difference of the classifiers on 8 datasets in Fig. 1, with a significance level of 0.05. It can be figured out that: The accuracy of AODSP is significantly higher, and the classification accuracy is greatly improved comparing to AODE and BCEP.

Table 3. Comparison among algorithms' error rate (%).

Data file	NB	AODE	ASAODE	BCEP	AODSP
balance-scale	**28.80**	30.40	29.44	54.24	30.11
bupa	36.81	36.81	36.81	42.03	36.81
german	24.66	**23.44**	23.80	25.50	24.22
house-votes-84	9.93	5.79	5.52	10.24	**5.20**
iris	5.60	6.80	6.00	33.33	**4.67**
labor	7.72	7.37	**5.26**	10.00	**5.26**
new-thyroid	**3.72**	4.19	5.58	30.23	4.47
pid	22.16	21.98	**21.22**	23.20	21.90
average	17.43	17.10	16.70	28.60	**16.58**

All classifiers except BCEP are part of a single clique, namely, most classifiers do not obtain significantly higher or lower results in terms of accuracy.

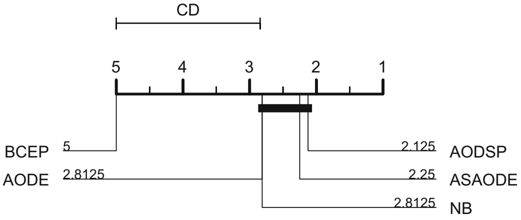

Fig. 1. Critical difference diagram for different classifiers on the 8 data sets.

5 Conclusion and Future Work

The naïve Bayesian classification is effective but restrictive due to its conditional independence assumptions. In order to weaken the conditional independence of the naïve Bayesian classification algorithm, in this paper, attributes selection is made based on the patterns composed of "attribute-value" pairs, and a lazy one-dependence classification algorithm based on selective patterns is proposed. The discrimination of selective patterns that play major roles as the basis for classification is mined, and the relationship between two types of attributes (belonging to selective patterns or not) is analyzed by the Bayesian network. The accuracy of the proposed algorithm is verified by using 8 datasets in the UCI machine learning database as experimental data. Based on the further analysis of the experimental results, it proves that the classification ability can be effectively improved by using the discrimination of patterns and dealing with dependencies between the attributes. In future work, it is necessary to further consider dependencies between the attributes and explore more appropriate patterns' selection criteria.

References

1. Domingos, P., Pazzani, M.: Beyond independence: conditions for the optimality of the simple Bayesian classifier. In: Saitta, L. (ed.) Proceedings of the 13th ICML, pp. 105–112. Morgan Kaufmann, San Francisco (1996)
2. Friedman, N., Goldszmidt, M.: Building classifiers using Bayesian networks. In: Proceedings of the 13th National Conference on Artificial Intelligence (AAAI 1996), pp. 1277–1284. AAAI Press, Menlo Park (1996)
3. Webb, G., Boughton, J., Wang, Z.: Not so naïve Bayes: aggregating one-dependence estimators. Mach. Learn. **58**(1), 5–24 (2005)
4. Chen, S., Martínez, A.M., Webb, G.I.: Highly scalable attribute selection for averaged one-dependence estimators. In: Tseng, V.S., Ho, T.B., Zhou, Z.-H., Chen, A.L.P., Kao, H.-Y. (eds.) PAKDD 2014. LNCS (LNAI), vol. 8444, pp. 86–97. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-06605-9_8
5. Chen, S., Martínez, A.M., Webb, G.I., Wang, L.: Selective AnDE for large data learning: a low-bias memory constrained approach. Knowl. Inf. Syst. **50**(2), 475–503 (2017)
6. Yu, L., Jiang, L., Wang, D., Zhang, L.: Attribute value weighted average of one-dependence estimators. Entropy **19**(9), 501 (2017)
7. Dong, G., Zhang, X., Wong, L., Li, J.: CAEP: classification by aggregating emerging patterns. In: Arikawa, S., Furukawa, K. (eds.) DS 1999. LNCS (LNAI), vol. 1721, pp. 30–42. Springer, Heidelberg (1999). https://doi.org/10.1007/3-540-46846-3_4
8. Fan, H., Ramamohanarao, K.: A Bayesian approach to use emerging patterns for classification. In: Schewe, K., Zhou, X. (eds.) Proceedings of the 14th Australasian Database Conference, pp. 39–48. ACS Press, Adelaide, Australia (2003)
9. Li, J., Dong, G., Ramamohanarao, K.: Making use of the most expressive jumping emerging patterns for classification. Knowl. Inf. Syst. **3**(2), 131–145 (2001)
10. Dong, G., Li, J.: Efficient mining of emerging patterns: discovering trends and differences. In: Proceedings of the 5th ACM SIGKDD International Conference on KDD, pp. 43–52. ACM Press, New York (1999)
11. Blake, C., Merz, C.: UCI repository of machine learning databases. http://archive.ics.uci.edu/ml/index.html. Accessed 1 June 2018
12. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuous valued attributes for classification learning. In: Bajcsy, R. (ed.) Proceedings of the 13th International Joint Conference on Artificial Intelligence, pp. 1022–1027. Morgan Kaufmann, San Mateo, CA (1993)
13. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**(1), 1–30 (2006)