A Review of Classification Algorithms for Data Mining

Li Mindong[1,a], Chen Qingwei[1,b], Huang Panling[1,c], Zhou Jun[1,d,*], Gong Weike[2]

[1]Shandong University, Jinan, China

[2]Shandong Bosheng Power Technology CO.,LTD, Linyi, China

[a]1102570384@qq.com, [b]858747715@qq.com, [c]hfpl@sdu.edu.cn, [d]zhoujun@sdu.edu.cn

Keywords: data mining; classification algorithm; review

Abstract: Classification algorithm is one of the important algorithms in data mining. Common classification algorithms such as decision tree, Bayesian network, support vector machine, association rules based classification algorithm and K-nearest neighbor algorithm have been widely used. This paper introduces the classical classification algorithm, compares the advantages and disadvantages of each algorithm and the latest research progress of each algorithm.

1. Introduction

Classification algorithm has always been one of the hotspots in data mining research. Classification algorithm discovers classification rules through the analysis of data training sets, and thus has the ability to predict new data types. Classification algorithms are evaluated by predictive accuracy, classification speed, robustness, scalability, and interpretability In this paper, according to the characteristics of various algorithms, they are divided into decision tree, Bayesian classification, support vector machine, classification of association rules and K-nearest neighbor and so on. It is convenient for developers and researchers to select and study classification algorithms.

2. Decision Tree

2.1 Algorithm Introduction

Decision tree is a basic classification and regression method. The decision tree model is a tree structure that describes the classification of instances, as shown in the following figure 1. Start from the root node, then test a feature of the instance, and assign the instance to its children according to the test results.

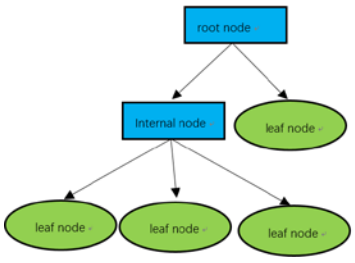

Fig. 1. Decision tree model diagram.

2.2 Algorithm Analysis

Decision tree classification is the process of classifying instances based on features. The model has the advantages of readability and fast classification speed. First, the decision tree model is established based on the training data and minimizing the loss function. Then, the new data is classified by decision tree model. Decision tree classification is widely applied. Azad et al. [1] Proposed an intrusion detection system based on decision tree and genetic algorithm. Wang [2]

applied decision tree algorithm to vehicle service, and proposed a new VHO method based on IOVS self-selection decision tree. The switching time is reduced, and the network update rate and vehicle service quality are guaranteed.

2.3 C4.5 Algorithm

The most widely used decision tree is the C4.5 algorithm. C4.5 algorithm is an improved version of decision tree algorithm ID3, which eliminates impossible branches and over-fitting branches, thus avoiding the over-fitting problem and greatly improving the calculation speed. The main advantages of C4.5 algorithm are high accuracy and fast construction of classification model, but it requires high quantity and quality of training samples. A Cherfi [3] proposed a new decision tree algorithm VFC4.5. Compared with C4.5 algorithm, VFC4.5 algorithm results in smaller decision tree in most cases and has better precision. L Chen [4] proposed a C4.5-K algorithm for futures data, which improved the ability of futures prediction.

3. Classification Based On Association

3.1 Algorithm Introduction

Classification based on association (CBA) [5] is a classification algorithm based on association rules discovery method. The algorithm consists of two steps to construct a classifier: the first step is to discover all classification association rules (CAR). The second step is to select high priority rules from the discovered CAR to overlay the training set. At present, there are many researches on this process, so the algorithm does not need to scan the training set too much in this step. Wang L [6] proposes a new algorithm based on quantitative association rules tree (CRQAR-tree), which combines association classification with rule-based TS fuzzy reasoning to generate rule tree structure for classification and regression prediction.

3.2 Algorithm Analysis

The discovery of association rules is based on classical algorithm Apriori. CBA rules is more accurate than decision tree C4.5 because it discovers relatively complete rules. CBA has high accuracy, but it is easy to be restricted by hardware memory. CBA is widely applied. Shao YX et al. [7] proposed a software defect prediction based on atomic class-association rule mining (ACAR). Alwidian J [8] used WCBA algorithm to propose a statistical measure based pruning and prediction technology, accurate prediction of breast cancer. Aljuboori [9] proposed a case-based reasoning association rule (CBRAR) strategy to improve the performance of similarity-based retrieval SBR classification frequent pattern tree FP-CAR algorithm, eliminating the ambiguity of case-based reasoning (CBR) error retrieval.

4. Bayes Classification

4.1 Algorithm Introduction

Bayes classification algorithm is a kind of classification algorithm based on probability and statistics, and the most widely used is the simple Bias classifier [10]. The description of the model is shown in Figure 2.

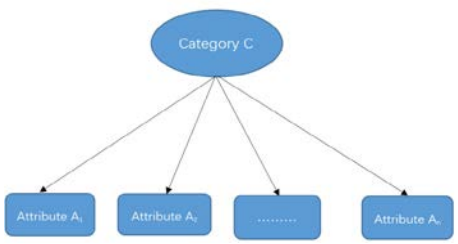

Fig. 2. naive Bayesian classification model.

Suppose there is a variable set $U=\{A,C\}$,among them,$A=\{A1,A2,\ldots An\}$ includes n conditional attributes. $C=\{C1,C2,\ldots Cn\}$contains m class labels. The naive Bayes classifier model assumes all the conditional attributes Ai (i=1,2,…,n)are all child nodes of class variable C. Assign a sample $X=\{a1,a2,\ldots,an\}$ to be classified to Ci ($1\leq i\leq m$), if and only if: $P(Ci\,|X)>P(Cj|X)(1\leq i,j\leq m,j\neq i)$.

According to Bayes theorem:

$$P(C_i \mid X) = \frac{(P(C_i)P(X|C_i)}{P(X)} \tag{1}$$

The steps of naive Bias classification model:

1) Preprocessing the required data sets, including discretization of attribute values and filling of missing values.

2) Simple size is S, the number of samples for class C_i is S_i, The number of samples in attribute A_k of class C_i is S_{ik}.

3) Calculate $P(C_i)=S_i/S$ and $P(a_k \mid C_i)=S_{ik}/S_i$.

4) According to the classification model, then get the decision result of the sample X to be classified.

4.2 Algorithm Analysis

The simple Bias classification algorithm has the following advantages:

1) Simple calculation.

2) Algorithm has higher stability. It is still effective when the amount of data is small.

3) Suitable for multi-class classification.

The disadvantage is that the independent hypothesis of attributes may not be well satisfied in practical problems, so the classification effect is difficult to achieve the actual expectations.

In recent years, as an important method of data classification, naive Bayes classification algorithm has been widely studied in the field of data mining because of its solid theoretical basis. Wang XG [11] proposes a weighted naive Bayesian algorithm based on MLRM (multiple linear regression model) to improve the classification accuracy and effectively improve the performance of NBC (naive Bayesian classification algorithm). L Zhao [12] proposes a conditional entropy matching semi-naive Bayesian classifier, which can effectively improve the performance of naive Bayesian classifier. Bayesian method is also used in image recognition. Ryan [13] proposes an optimized naive Bayesian algorithm and its application in face recognition.

5. K-Nearest Neighbor

5.1 Algorithm Introduction

K-Nearest Neighbor(KNN)'s main idea is to find the training set with the most similar features, so the type of sample to be predicted is the type of the nearest neighbor sample. The implementation of KNN algorithm is as follows:

1) Calculate the distance between feature vectors and feature vectors of each training set. Simple calculation.

2) Sort training samples by distance.

3) Take the first K samples in order, and count the labels of the samples with the most occurrences.

4) The highest frequency tag is considered to be the label of the sample to be predicted.

5.2 Algorithm Analysis

The classification decision rules in K-nearest neighbor algorithm are usually majority votes, and the class of input instance is determined by the majority of the classes in the K-nearest neighbor training instances of the input instance. Most voting rules refer to: if the classification loss function is a 0-1 loss function, the classification function is:

$$f: R^n \rightarrow \{c_1, c_2, \ldots, c_k\} \tag{2}$$

The probability of misclassification is:

$$P(Y \neq f(X)) = 1 - P(Y = f(X)) \tag{3}$$

For a given instance $x \in \chi$, its nearest neighbor's K training instance points constitute the set $N_k(x)$. If the category covering N_k (x) is c_j, then the misclassification rate is:

$$\frac{1}{k}\sum_{x_i \in N_k(x)} I(y_i \neq c_i) = 1 - \frac{1}{k}\sum_{x_i \in N_k(x)} I(y_i = c_i) \tag{4}$$

The simple Bias classification algorithm has the following advantages:

1) The algorithm is simple.

2) It can effectively avoid the imbalance of sample size. Simple calculation.

3) The accuracy of measurement is higher.

However, each sample of the prediction set needs to calculate its similarity with each training sample. The computational complexity is large, especially when the training set is large, the computational complexity will seriously affect the performance of the algorithm.

Because of the high accuracy of KNN algorithm, researchers have done many researches on K-nearest neighbor algorithm. Zeng Y [14] proposed the IML-KNN algorithm, which considers the influence of the nearest neighbors and K neighbors of non-classified samples on the basis of the traditional ML-KNN algorithm. It has a good classification effect for multi label evaluation indexes. Xie Y [15] proposes an improved KNN method, called KNN++, to classify complex data with heterogeneous views. Kutylowska M [16] uses nonparametric regression algorithm K nearest neighbor to predict failure rate. Zhang NA [17] proposes an improved KNN algorithm to overcome class overlapping problem when class distribution is skewed.

6. Support Vector Machine

6.1 Algorithm Introduction

Support Vector Machine (SVM) [18] is an algorithm developed from the optimal classification surface under linear separable condition. It has a strong theoretical basis and can be used in classification tasks. The equation of hyperplane is:

$$w^T x + b = 0 \tag{5}$$

$w = (w_1, w_2 ... w_k)$ is the normal vector of the hyperplane, representing the direction of the plane, b is the displacement, representing the distance between the hyperplane and the origin. Then we can get the distance from the point to the hyperplane in the sample is:

$$r = \frac{|w^T x + b|}{||w||} \tag{6}$$

If the hyperplane can separate the positive and negative samples, then we'll get:

$$\begin{cases} w^T x_i + b \geq +1, & y_i = +1 \\ w^T x_i + b \leq -1, & y_i = -1 \end{cases} \tag{7}$$

Distance between two support vectors from different types to hyperplanes is:

$$R = \frac{2}{||w||} \tag{8}$$

We hope that R will be the largest, so we need to get the following formula:

$$\min \frac{1}{2}||w||^2 \tag{9}$$

Lagrange equation can be obtained by adding Lagrange multiplier ∂_i≥0, we can get Lagrange equation:

$$L(\omega, b, \partial) = \frac{1}{2}||\omega||^2 + \sum_{i=1}^{k} \partial_i (1 - y_i(\omega^T x_i + b)) \tag{10}$$

So the SVM model is:

$$f(x) = \omega^T x + b = \sum_{i=1}^{k} \partial_i y_i x_i^T x + b \tag{11}$$

6.2 Algorithm Analysis

The complexity of SVM algorithm depends on the number of support vectors, not the dimension of sample space, so the computation is not very large and the generalization accuracy is high. The disadvantage is that the support vector algorithm is sensitive to parameter adjustment and kernel function selection, and it occupies more memory and running time in storage and calculation, so it is inadequate in large-scale sample training. HW Wang [19] effectively improves the accuracy of distribution using improved support vector machines. Yeh JP et al. [20] Proposed simulated annealing algorithm to reduce the number of support vector machines, and improve the classification accuracy. F Zhu [21] proposed a weighted class of support vector machines (WOC-SVM), which minimizes the impact of noise by assigning lower weights.

7. Other Classification Algorithms

In addition to the above classification algorithms, there are commonly used genetic algorithm, neural network and other classification algorithms. Genetic algorithm is one of the key technologies in modern intelligent computing, which is based on the idea of biological evolution and searches the optimal solution by simulating the natural evolution process. The neural network is a set of connected I/O units, where each connection is associated with a weight. Neural network algorithm has high tolerance to noise data and high classification accuracy, which makes neural network have a good effect in data mining. But it takes a lot of time to process large amounts of data.

8. Summary and Prospect

Classification is an important data mining technology. In this paper, various algorithms are summarized, and conclude the latest developments and application areas of each algorithm. In fact, in the era of expanding data, the performance of the algorithm is more important, such as speed of execution, scalability and comprehensibility of the output results. Therefore, although each algorithm has its own advantages, but a classification algorithm with good characteristics in all aspects is still worthy of further study.

Acknowledgments

1) Acknowledgments: This project is supported by Key R & D project of Shandong Province (Grant No. 2017CXGC0810).

2) Acknowledgments: This project is supported by Key R & D project of Shandong Province (Grant No. 2017CXGC0215).

3) Acknowledgments: This project is supported by Key R & D project of Shandong Province (Grant No. 2017CXGC0903).

4) Acknowledgments: This project is supported by Key R & D project of Shandong Province (Grant No. 2018CXGC0908).

5) Acknowledgments: This project is supported by Key R & D project of Shandong Province (Grant No. 2018CXGC0215).

6) Acknowledgments: This project is supported by Key R & D project of Shandong Province (Grant No. 2018CXGC1405).

7） Acknowledgments: This project is supported by Key R & D project of Shandong Province

(Grant No. 2018CXGC0808).

8) Acknowledgments: This project is supported by Key R & D project of Shandong Province (Grant No. 2018CXGC0601).

References

[1] V. Nath and J. Kumar Mandal, “Decision tree and genetic algorithm based intrusion detection system,” Second International Conference on Microelectronics, Computing & Communication Systems，Ranchi, pp.141-52, May 2017.

[2] SG. Wang, CQ. Fan, CH. Hsu, QB. Sun and FC. Yang, “A Vertical Handoff Method via Self-Selection Decision Tree for Internet of Vehicles”, IEEE SYSTEMS JOURNAL, vol. 10, pp. 1183-1192, Sep 2016.

[3] A. Cherfi, K. Nouira and A. Ferchichi, “Very Fast C4.5 Decision Tree Algorithm,” Applied Artificial Intelligence, vol.32, pp. 110-137, 2018.

[4] Chen L and Guohui H E, “Research on improved C4.5 algorithm in futures data mining,” Computer Engineering & Applications, vol. 53, pp. 161-166, 2017.

[5] B. Liu, W. Hsu, Y. Ma, “Integrating classification and association rule mining,” International Conference on Knowledge Discovery and Data Mining, vol. 1711, pp. 80-86, 1998.

[6] L. Wang, SL. Li, H. Sun, KX. Peng, “A classification and regression algorithm based on quantitative association rule tree,” JOURNAL OF INTELLIGENT & FUZZY SYSTEMS, vol. 31, pp. 1407-1418, 2016.

[7] YX. Shao, B. Liu, SH. Wang and GQ. Li, “A novel software defect prediction based on atomic class-association rule mining,” Expert system with applications, vol. 114, pp. 237-254, 2018.

[8] J. Alwidian, BH. Hammo, N. Obeid, “WCBA: Weighted classification based on association rules algorithm for breast cancer disease,” Applied soft computer, vol. 62, pp. 536-549, 2018.

[9] Aljuboori A, Meziane F, Parsons D. A New Strategy for Case-Based Reasoning Retrieval Using Classification Based on Association[J]. 2016.

[10] Nir Friedman, Dan Geiger, Moises Goldszmidt. Bayesian Network Classifiers[J]. Machine Learning, 1997, 29(2-3):131-163.

[11] XG. Wang, X. Sun, “An improved weighted naive Bayesian classification algorithm based on multivariable linear regression model,” ISCID, vol.2 , pp. 219-222, 2016.

[12] Zhao L, Liu J, Cui C, et al. Semi-naive Bayesian classifier matched by mutual information[J]. Computer Engineering & Applications, 2016.

[13] Yan R, Wen J, Cao J, et al. An Optimized Naive Bayesian Method for Face Recognition[C]// International Conference on Cognitive Systems and Signal Processing. Springer, Singapore, 2016:126-135.

[14] Y. Zeng, HM. Fu, YP. Zhang, XY. Zhao, “An Improved ML-kNN Algorithm by Fusing Nearest Neighbor Classification,” AICS, pp. 193-198, 216.

[15] Y. Xie, “KNN plus plus : An Enhanced K-Nearest Neighbor Approach for Classifying Data with Heterogeneous Views,” Hybrid intelligent systems, vol. 420, pp. 13-23, 2016.

[16] M. Kutylowska, “K-Nearest Neighbours Method as a Tool for Failure Rate Prediction,”Periodica polytechnica-civil engineering, vol.62, pp. 318-322, 2018.

[17] NA. Zhang, W. Karimoune, L. Thompson, HM. Dang, “A Between-Class Overlapping Coherence-Based Algorithm in KNN Classification,” IEEE International Conference on Systems, Banff, Canada, pp. 572-577, 2017.

[18] Vapnik V, Levin E, Cun Y L. Measuring the VC-dimension of a learning machine[J]. Neural Computation, 1994, 6(5):851-876.

[19] Wang H W, Kong B. An Improved Weighted Support Vector Machine[J]. Journal of Henan Normal University, 2011, 39(3):167-170.

[20] Yeh J P, Chiang C M. Reducing the Solution of Support Vector Machines Using Simulated Annealing Algorithm[C]// International Conference on Control, Artificial Intelligence, Robotics & Optimization. IEEE Computer Society, 2017:105-108.

[21] Zhu F, Yang J, Gao C, et al. A weighted one-class support vector machine[J]. Neurocomputing, 2016, 189:1-10.