Research Journal of Applied Sciences, Engineering and Technology 5(4): 1420-1423, 2013
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: July 12, 2012 Accepted: August 15, 2012 Published: February 01, 2013

Classification Based on Attribute Positive Correlation and Average Similarity of Nearest Neighbors

Zhongmei Zhou, Guiying Pan and Xuejun Wang
Department of Computer Science and Engineering, Zhangzhou Normal University, Zhangzhou, 363000, China

Abstract: The K-Nearest Neighbor algorithm (KNN) is a method for classifying objects based on the k closest training objects. An object is classified by a majority vote of its nearest neighbors. "Closeness" is defined in terms of the similarity measure between two objects. KNN is not only simple, but also sometimes has high accuracy. However, the quality of KNN classification result depends on the similarity measure between two objects and the selection of k. Moreover, the average similarity of the majority nearest neighbors may be less than the one of the minority nearest neighbors. To deal with these problems, in this study, we propose a new classification approach called APCAS: classification based on the attribute values which are positively correlated with one of the class labels and the average similarity of the nearest neighbors in each class. First, we define a new similarity measure based on the attribute values which are positively correlated with one of the class labels. Second, we classify a new object using the average similarity of the nearest neighbors in each class without selecting k. Experimental results on the mushroom data show that APCAS achieves high accuracy.

Keywords: KNN, nearest neighbor, positive correlation, similarity measure

INTRODUCTION

KNN is widely discussed and applied in pattern recognition (Zhang *et al*., 2004). KNN is a method for classifying objects based on the k closest training examples. "Closeness" is defined in terms of the similarity measure between two objects. KNN is among the simplest of all machine learning algorithms. However, firstly, the classification accuracy of KNN depends on the similarity measure between two objects. In order to achieve high accuracy, (Feng *et al*., 2005) proposed KNN-M algorithm for text categorization. The major difference between KNN-M and KNN lies in the calculation of text similarity on finding k-nearest neighbors. With categorical variable, the similarity between two objects is often computed using the simple matching approach. Nevertheless, the simple attribute value matching cannot reflect the importance of these matched attribute values to the class label. If two objects are similar, then they not only have many same attribute values and also have many same attribute values which are important to one of the class labels. If an attribute value is important, then it must be positively correlated with one of the class labels.

To deal with the problem of similarity measure, we define a new similarity measure based on the attribute values which are positively correlated with one of the class labels. If two objects are similar, then they have many same attribute values which are positively correlated with one of the class labels. A difficulty in this study is that there are few correlation measures which have proper bounds for effectively evaluating the correlation degree between the attribute value and the class label. The most commonly employed method for correlation mining is that of two-dimensional contingency table analysis of categorical data using the chi-square statistic as a measure of significance. Brin *et al*. (1997) analyzed contingency tables to generate correlation rules that identify statistical correlation in both the presence and absence of items in patterns. Liu *et al*. (1999) analyzed contingency tables to discover unexpected and interesting patterns that have low level of support and high level of confidence. Bing *et al*. (1999) used contingency tables for pruning and discovered correlations etc. Although the low chi-squared value (less than the cutoff value, e.g., 3.84 at the 95% significance lever) effectively indicates that all patterns AB, $\bar{A}$B, A$\bar{B}$, $\bar{A}\bar{B}$ are independent, that is, A and B, $\bar{A}$ and B, A and $\bar{B}$, $\bar{A}$ and $\bar{B}$ are all independent. The high chi-squared value only indicates that at least one of patterns $AB, \bar{A}B, A\bar{B}, \bar{A}\bar{B}$ is not independent, so it is possible that A and B are independent, in spite of the high chi-squared value. Therefore, the chi-squared value is not reasonable for measuring the correlation degree of A and B.

Corresponding Author: Zhongmei Zhou, Department of Computer Science and Engineering, Zhangzhou Normal University, Zhangzhou 363000, China

For other commonly used measures, the measure P(AB)/P(A)P(B) does not have proper bounds. P(AB)-P(A)P(B) (Piatetsky-Shapiro, 1991) is not rational when P(AB) is compared with P(A)P(B). For example, if P(AB) = 0.02, P(A)P(B) = 0.01, P(CD) = 0.99 and P(C)P(D) = 0.98, P(AB)-P(A)P(B) = P(CD)-P(C)P(D). The correlation degree of A and B is equal to the correlation degree of C and D. But, P(AB)/P(A) P(B) = 2 and P(CD)/P(C)P(D) = 1.01, the correlation degree of A and B is much higher than the correlation degree of C and D. In this study, we use the correlation measure correlation confidence (Zhong *et al*., 2006) to evaluate the correlation between two items. The measure correlation confidence has two bounds -1 and 1. We can see from Zhong *et al*. (2006) that the measure correlation confidence is reasonable.

Secondly, the quality of KNN classification result depends on the selection of k. The best choice of k depends upon the data. It is difficult to select an appropriate k (Anil, 2006). Moreover, the average similarity of the majority nearest neighbors may be less than the one of the minority nearest neighbors. Thus, it may be unreasonable to classify a new object by a majority vote. Therefore, we propose a new classification approach called APCAS. We not only use a new similarity measure and also classify a new object using the average similarity of the nearest neighbors in each class without selecting k. Experimental results on the mushroom data set show that APCAS achieves high accuracy.

METHOD AND EXAMPLE

In this section, we first introduce some related definitions and then give an example to explain the classification algorithm APCAS.

In statistical theory, $X_1, X_2,\ldots, X_n$ are independent if and only if $\forall k$ and $\forall 1 \le i_1 < i_2 < \cdots < i_k \le n$:

$$P(X_{i1}X_{i2}\ldots X_{ik}) = P(X_{i1})P(X_{i2})\ldots.P(X_{ik}) \quad (1)$$

We use the correlation measure correlation confidence to evaluate the degree of correlation relationships between any two objects. The correlation confidence of any two objects C and D is defined as follows (Zhong *et al*., 2006):

$$\rho(CD) = P(CD) - P(C)P(D) / P(CD) + P(C)P(D) \quad (2)$$

From (2), we can see that P(CD) has two bounds -1 and 1. According to the conception of correlation in statistical theory, any two objects C and D are positively correlated if and only if P(CD)>0.

The training data set T has m distinct attributes A1, A2,… Am and a list of classes C1, C2,… Cm. All attribute values are categorical. We define a new similarity measure between two objects as seen from definition 2:

Table 1: A data set

id	*A*	*B*	*C*	*D*	Class
X1	32	55	80	83	90
X2	33	52	80	85	89
X3	33	55	79	88	90
X4	34	55	79	82	89
X5	32	55	80	88	90
X6	34	55	77	82	89
X7	33	55	79	82	90

Definition 1: (Similarity 1) X and Y are two object. If X and Y have same attribute values v1, v2,… vm, then the similarity of X and Y is n , i.e., s1(x, y) = n.

Definition 2: (Similarity 2) X and Y are two object. Let X and Y have same attribute values v1, v2,… vm,. P(vi) Is the probability of attribute value vi.cj. Is a class label. If Y is belonging to cj and P(vicj)>0, then the similarity of X and Y is defined as follows:

$$S2(x, y) = \sum_{i=1}^{n} P(vi)\rho(vicj) \quad (3)$$

From definition 2, we can see that if X and Y have high similarity, then they must have same attribute values which are positively correlated with one of the class labels and at the same time have high probability.

We illustrate the classification algorithm APCAS using the following example.

Example 1: Given a data set T as shown in Table 1. X7 is a test object.

By similarity 1, we have s1(x7, x1) = 1, s1(x7, x2) = 1, s1(x7, x3) = 3, s1(x7, x4) = 3, s1(x7, x5) = 1 and s1(x7, x6) = 2. The average similarity $\overline{S1}$ of all training objects is 11/6. X7 has nearest neighbors X4and X6 in class 89. X7 has a nearest neighbor X3 in class 90. The average similarity of the nearest neighbors in class 89 is 25. The average similarity of the nearest neighbors in class 90 is 3. Therefore, we assign the test object X7 to the class 90.

The classification algorithm APCAS use the similarity measure by definition 2. We have:

P(33) = 1/3, P(55) = 5/6, P(79) = 1/3, P(82) = 1/3,

P(33, 89) = P(33, 90) 0, P(79, 89) = P(79, 90) = 0,

P(82, 89) = P (82, 89) = 0, P(55, 90) = 1/11

So, the attribute value 55 is positively correlated with the class 90. By definition 2, we have:

s2(x7, x1) = s1(x7, x3) = s1(x7, x5) = 5/66, s2(x7, x2) = s1(x7, x4) = s1(x7, x6) = 0

The average similarity $\bar{s}2$ of all training objects is 11/132. X7 has no nearest neighbors in class 89. X7 has nearest neighbors X1, X3 and X5 in class 90. The average similarity of the nearest neighbors in class 89 is 0. The average similarity of the nearest neighbors in class 90 is 5/66. So, we assign the test object X7 to the class 90.

From the example, we can see that it is reasonable for us to use the average similarity of nearest neighbors in each class.

EXPERIMENTAL RESULTS

All experiments are performed on mushroom characteristic dataset, which consists of 5643 objects. All objects have 23 attribute values.

We classify a new object using the average similarity of all training objects in each class in the first algorithm A1 and the third algorithm A3. We classify a new object using the average similarity of its nearest neighbors in each class in the second algorithm A2 and the forth algorithm APCAS. In both algorithm A1 and algorithm A2, we use similarity measure defined in definition 1. In both algorithm A3 and algorithm APCAS, we use similarity measure defined in definition 2.

In Table 2, we select training set by random named T-set. We select training objects in turn from 100 to 500. We select every 500 objects as test set in turn from 1 to 5000. We compare the average classification accuracy of algorithm APCAS with algorithm A1, A2 and A3. From Table 2, we can see that algorithm APCAS have higher classification accuracy than other algorithms. From Table 2, we can also see that algorithm A2 have higher classification accuracy than algorithms A1. Therefore, we can conclude from Table 2 that:

- It is reasonable to use the average similarity of the nearest neighbors in each class.
- Similarity measure defined in definition 2 is better than the one defined in definition 1.

In Table 3, we select 100 training objects by random. We select every 500 objects as test set in turn from 1 to 5000. We compare the classification accuracy of algorithm APCAS with algorithm A1, A2 and A3. From Table 3, we can see that algorithm APCAS have higher accuracy than other algorithms in every time.

In Table 4, we select 500 training objects by random. We select every 500 objects as test set in turn from 1 to 5000. We compare the classification accuracy of algorithm APCAS with algorithm A1, A2 and A3.

Table 2: The comparison on average accuracy

Tr-set	*A1*	*A2*	*A3*	APCAS
100	0.895	0.972	0.985	0.989
200	0.905	0.970	0.968	0.972
300	0.904	0.975	0.980	0.987
400	0.892	0.970	0.963	0.975
500	0.910	0.970	0.965	0.977

Table 3: The comparison on accuracy (100)

Te-set	*A1*	*A2*	*A3*	APCAS
500	0.918	0.966	0.978	0.988
500	0.880	0.984	0.994	0.994
500	0.914	0.968	0.982	0.982
500	0.894	0.970	0.978	0.982
500	0.892	0.972	0.984	0.988
500	0.878	0.966	0.982	0.986
500	0.888	0.970	0.986	0.990
500	0.888	0.968	0.992	0.996
500	0.904	0.980	0.990	0.992

Table 4: The comparison on accuracy (500)

Te-set	*A1*	*A2*	*A3*	APCAS
500	0.926	0.956	0.962	0.978
500	0.934	0.970	0.968	0.984
500	0.926	0.978	0.962	0.972
500	0.914	0.976	0.956	0.970
500	0.894	0.974	0.966	0.980
500	0.894	0.968	0.958	0.974
500	0.894	0.966	0.960	0.974
500	0.898	0.964	0.972	0.978
500	0.910	0.980	0.970	0.980

From Table 4, we can see that algorithm APCAS have higher accuracy than other algorithms in many times.

CONCLUSION

Although KNN is simple, it suffers from some deficiencies. In this study, we proposed a new classification algorithm APCAS based on a new similarity measure. While measuring similarity between two objects, we not only think about the importance of an attribute value to the class label, but also consider the number of matched attribute values. If an attribute value is important, then it is not only positively correlated with one of the class labels, but also has high probability. If two objects are similarity, then they must have same attribute values which are important to the class label. In order to achieve high classification accuracy, we classify a new object by the average similarity of its nearest neighbors in each class. Experimental results show that APCAS achieves higher classification accuracy.

ACKNOWLEDGMENT

This study is supported in part by China NSF program (No. 61170129, 10971186), a grant from education ministry of Fujian, China (No. JA10202).

REFERENCES

Anil, K.G., 2006. Optimum choice of k in nearest neighbor classification. Comput. Stat. Data An., 50: 3113-3123.

Bing, L., H. Wynne and M. Yiming, 1999. Pruning and summarizing the discovered associations. Proceeding of ACM SIGKDD, International Conference of Knowledge Discovery in Databases, New York, USA, pp: 125-134

Brin, S., R. Motwani and C. Silverstein, 1997. Beyond market baskets: Generalizing association rules to correlations. Proceeding of ACM SIGMOD Int. Conf. Manageme. Data, 26(2): 265-276.

Feng, Y., H.W. Zhao and M.Z. Zhong, 2005. Combining an order-semisensitive text similarity and closest fit approach to textual missing values in knowledge discovery. KES (2), pp: 943-949.

Liu, H., H. Lu, L. Feng and F. Hussain, 1999. Efficient search of reliable exceptions. Proceeding of Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer-Verlag, London, UK, pp: 194-203.

Piatetsky-Shapiro, G., 1991. Discovery, Analysis and Presentation of Strong Rules. Knowledge Discovery in Databases, AAAI/MIT Press, pp: 229-248.

Zhong, M.Z., H.W. Zhao, S.W. Chun and Y. Feng, 2006. Efficiently mining mutually and positively correlated patterns. ADMA, LNAI, 4093: 118-125.

Zhang, B. and S.N. Srihari, 2004. Fast k-nearest neighbors classification using cluster-based trees. IEEE T. Pattern Anal., 26: 525-528.