Diversity Measure for Multiple Classifier Systems

Qinghua Hu and Daren Yu

Harbin Institute of Technology, Harbin, China
Huqinghua@hcms.hit.edu.cn

Abstract. Multiple classifier systems have become a popular classification paradigm for strong generalization performance. Diversity measures play an important role in constructing and explaining multiple classifier systems. A diversity measure based on relation entropy is proposed in this paper. The entropy will increase with diversity in ensembles. We introduce a technique to build rough decision forests, which selectively combine some decision trees trained with multiple reducts of the original data based on the simple genetic algorithm. Experiments show that selective multiple classifier systems with genetic algorithms get greater entropy than those of the top-classifier systems. Accordingly, good performance is consistently derived from the GA based multiple classifier systems although accuracies of individuals are weak relative to top-classifier systems, which shows the proposed relation entropy is a consistent diversity measure for multiple classifier systems.

1 Introduction

In the last decade, multiple classifier systems (MCS) become a popular technique for building a pattern recognition machine [4, 5]. This system is to construct several distinct classifiers, and then combines their predictions. It has been observed the objects misclassified by one classifier would not necessarily misclassified by another, which suggests that different classifiers potentially offered complementary information. This paradigm is with several names in different views, such as neural network ensemble, committee machine, and decision forest. In order to construct a multiple classifier system, some techniques were exploited. The most widely used one is resampling, which selects a subset of training data with different algorithms. Resampling can be roughly grouped into two classes; one is to generate a series of training sets from the original training set and then trains a classifier with each subset. The second method is to use different feature sets in training classifiers. Random subspace method, feature selection were reported in the documents [2, 4].

The performance of multiple classifier systems not only depends on the power of the individual classifiers in the system, but also is influenced by the independence between individuals [5, 6]. Diversity plays an important role in combining multiple classifiers, which guilds MCS users to design a good ensemble and explain the success of a ensemble systems. Diversity may be interpreted differently from some angles, such as independence, orthogonality or complementarity [7, 8]. Kuncheva pointed that diversity is generally beneficial but it is not a substitute for accuracy [6].

L. Wang and Y. Jin (Eds.): FSKD 2005, LNAI 3613, pp. 1261 – 1265, 2005.

As there are some pair-wise measures, which cannot reflex the whole diversity in MCS, A novel diversity measure for the whole system is presented in the paper, called relation entropy, which is based on the pair-wise measures.

2 Relation Entropy

Here we firstly introduce two classical pairwise diversity measures, Q-statistic and correlation coefficient. Given a multiple classifier system with n individual classifiers $\{C_1, C_2, C_i, \cdots, C_n\}$, the joint output of two classifiers, C_i and C_j, $1 \le i, j \le n$, can be represented in a 2×2 table as shown in table 1.

Table 1. The relation table with classifiers C_i and C_j

	C_j correct (1)	C_j wrong (0)
C_i correct (1)	N^{11}	N^{10}
C_i wrong (0)	N^{01}	N^{00}

Yule introduced Q-statistic for two classifiers defined as

$$Q_{ij} = \frac{N^{11}N^{00} - N^{10}N^{01}}{N^{11}N^{00} + N^{10}N^{01}}$$

The correlation coefficient ρ_{ij} is defined as

$$\rho_{ij} = \frac{N^{11}N^{00} - N^{10}N^{01}}{\sqrt{(N^{11} + N^{10})(N^{01} + N^{00})(N^{11} + N^{01})(N^{10} + N^{00})}}$$

Compute the Q-statistic or correlation coefficient of each pair of n classifiers, a matrix will produce: $M = \left(r_{ij}\right)_{n \times n}$ Here $r_{ii} = 1$, $r_{ij} = r_{ji}$ and $|r_{ij}| \le 1$. Therefore matrix |*M*| is a fuzzy similarity relation matrix. the greater the value $|r_{ij}|$, $i \ne j$, is, the stronger the relation between C_i and C_j is and then the weaker of independence between classifiers is. The matrix *M* surveys the total relation of classifiers in the MCS.

Given a set of classifiers $C = \{C_1, C_2, \cdots, C_n\}$, R is a fuzzy relation on *C*. It can be denoted as a relation matrix $(R_{ij})_{n \times n}$, where R_{ij} is the relation degree between C_i and C_j with respect to relation R. As we know that the larger R_{ij} is, the stronger the relation of C_i and C_j is. As to correlation coefficient, R_{ij} denotes the degree of correlation between C_i and C_j. If $R_{ij} > R_{ik}$, we say C_i and C_j are more indiscernible than C_i and C_k.

Definition 1. Let R be a fuzzy relation over a set C, w_i the weight of C_i in the ensemble system, $0 \le w_i \le 1$ and $\sum_i w_i = 1$. $\forall C_i \in C$. We define expected relation degree of C_i to all $C_j \in C$ with respect to R as follows:

$$\pi(C_i) = \sum_{j=1}^{n} w_j \bullet r_{ij}$$

Definition 2. The information quantity of relation degree of C_i is defined as

$$I(C_i) = -\log_2 \pi(C_i)$$

It's easy to show that the larger $\pi(C_i)$ is, the stronger C_i is with other classifiers in the ensemble system, and the less $I(C_i)$ is, which shows that the measure $I(C_i)$ describes the relation degree of C_i to all classifiers in system C with respect to relation R.

Definition 3. Given any relation R between individuals in multiple classifier system, and a weight factor series of C, the relation entropy of the pair $<R, w>$ is defined as

$$H_w(R) = \sum_{C_i \in C} w_i \bullet I(C_i) = -\sum_{C_i \in C} w_i \log_2 \pi(C_i)$$

Information entropy gives the total diversity of a multiple classifier system if relations used represent the similarity of outputs of individual classifiers. This measure not only takes the relations between classifiers into account, but also computes the weight factors of individual classifiers in ensemble. The proposed information entropy can applied to a number of pairwise similarity measures for multiple classifier systems, such as Q-statistic, correlation coefficient and so on.

3 Experiments

Searching the optimal ensemble of multiple classifier systems involves combinational optimization. Genetic algorithms make a good performance in this kind of problems. Some experiments were conducted with UCI data. The numbers of reducts range between 5 and 229. All the trees are trained with CART algorithm and two-thirds samples in each class are selected as training set, others are test set. Here, for simplicity, 20 reducts are randomly extracted from the reduct sets of all data sets if there are more than 20 reducts. Subsequent experiments are conducted on the 20 reducts.

The accuracies with different decision forests are shown in table 2. GAS means the forests based on genetic algorithm. TOP denotes the forests with the best trees. We find that GAS ensembles get consistent improvement for all data sets relative to systems combining the best classifiers.

All entropies of Q-statistic and correlation coefficient in two kinds of ensembles as to the data sets are shown in table 3. As the entropies represent the total diversity in systems, we can find GAS based ensembles consistently catch more diversity than top-classifier ensembles.

Table 2. Comparison of decision forests

	GAS		TOP	
Data	size	accuracy	size	accuracy
BCW	10	0.9766	10	0.92642
Heart	6	0.8857	6	0.85714
Ionos	8	0.9901	8	0.94059
WDBC	7	0.9704	7	0.94675
Wine	7	1.00	7	0.97917
WPBC	9	0.75	9	0.70588

Table 3. Relation entropy of multiple classifier systems

	Q-statistic		**Correlation coefficient**	
	TOP	**GAS**	**TOP**	**GAS**
BCW	0.1385	0.2252	0.6348	0.7719
Heart	0.0031	0.4011	0.0698	0.9319
IONOS	0.0978	0.2313	0.7689	1.2639
WDBC	0.1780	0.2340	1.0296	1.2231
Wine	0	0.0593	0.8740	1.3399
WPBC	1.0541	1.1466	1.7425	1.8319

4 Conclusion

Diversity in multiple classifier systems plays an important role in improve classification accuracy and robustness as the performance of ensembles not only depends on the power of individuals in systems, but also is influenced by the independence between individuals. Diversity measures can guild users to select classifiers and explain the success of the multiple classifier system. Here a total diversity measure for multiple classifier systems is proposed in the paper. The measure computes the information entropy represented with a relation matrix. If Q-statistic or correlation coefficient is employed, the information quantity reflexes the diversity of the individuals. We compare two kinds of rough decision forest based multiple classifier systems with 9 UCI data sets. GA based selective ensembles achieve consistent improvement for all tasks compared with the ensembles with best classifiers. Correspondingly, we find the diversity of GAS with the proposed entropy based on Q-statistic and correlation coefficient is consistently greater than that of top-classifier ensembles, which shows that the proposed entropy can be used to explain the advantage of GA based ensembles.

References

1. Ghosh J.: Multiclassifier systems: Back to the future. Multiple classifier systems. Lecture notes in computer science, Vol.2364. Springer-Verlag, Berlin Heidelberg (2002) 1-15
2. Zhou Z., Wu J., Tang W.: Ensembling neural networks: Many could be better than all. Artificial intelligence 137 (2002) 239-263

3. Ho Ti Kam: Random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence. 20, (1998) 8, 832-844
4. Czyz J., Kittler J., Vandendorpe L.: Multiple classifier combination for face-based identity verification. Pattern recognition. 37 (2004) 7: 1459-1469
5. Ludmila I. Kuncheval: Diversity in multiple classifier systems. Information fusion. 6, (2005) 3-4
6. Kuncheva L. I. et al.: An experimental study on diversity for bagging and boosting with linear classifiers. Information fusion. 3 (2002) 245-258
7. Hu Qinghua, Yu Daren: Entropies of fuzzy indiscernibility relation and its operations. International journal of uncertainty, fuzziness and knowledge based systems. 12 (2004) 575-589
8. Hu Qinghua, Yu Daren, Wang Mingyang: Constructing rough decision forests. The tenth conference on rough sets, fuzzy sets, data mining and granular computing. 2005