Ensemble Learning Based on Parametric Triangular Norms

Pengtao Jia*

School of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an, Shaanxi, 710054, China

Abstract: Along with the increase of data usage in actual applications, it has become an important issue in ensemble learning to improve the ability for data analysis and processing. In order to improve the learning precision and get more accurate classification and projections in practical problems, triangular norms are introduced into the ensemble learning system. Triangular norms can improve the generalization capability of machine learning models. This paper investigates the effect of applying six different parametric triangular norms in ensemble learning system. Firstly, a new combination model for ensemble learning was constructed. Then, six different parametric triangular norms were used respectively as the combination rule of the new model. Finally, genetic algorithm was used as the parameter estimation module of the new rules. There are two kinds of experiments were conducted, they are the classification experiments and the prediction experiments. The classification experiments were conducted on seven different datasets from the University of California, Irvine machine learning repository (UCI). The prediction experiments were conducted on five samples with actual values. The experimental results show that choosing the appropriate combination rule may lead to higher accuracy than the single classifiers or the single prediction models. The improved performance is produced by employing the Yager operator, Aczel-Alsina operator and Schweizer-Sklar operator, and the Yager operator is most suitable for the combination rule of ensemble learning system.

Keywords: Combination model, ensemble learning, generalization capability, genetic algorithm, parametric triangular norms.

1. INTRODUCTION

In recent years, ensemble learning has received more and more attention of researchers from various fields, such as statistics, machine learning, pattern recognition and knowledge discovery. In machine learning, ensemble learning means to construct a set of hypotheses and combine them to use [1]. Typically, the generalization ability of an ensemble is much stronger than that of the base learners in it, so it can produce improved accuracy compared to a single learner or regression mode [2].

An ensemble can be constructed in two steps: produce base learners and then combine the base learners to use. The base learners should be adequately accurate and diverse, so as to make a good ensemble. Researches show that there have already some combination methods used to combine the results of base learners, such as the product rule, mean rule, median rule, max rule, min rule, the majority voting method (sample voting) [3], the weighted voting method [4], the stacked generalization [5], and the method by combining the nearest neighbor classifiers through multiple feature subsets [6]. However, in many practical applications, these methods do not perform better in different datasets. So, a new combination rule should be constructed to deal with different datasets and combine the results of base learners effectively.

Triangular norms are important family operations with strong generalization capability in fuzzy logic. References [7] introduced them to construct the fuzzy rule based classification systems and altered the overall accuracy of the systems. So, in this study, they are introduced to improve the performance of ensemble learning algorithm.

This study presents combination rule based on parametric triangular norms for ensemble learning. Consider the test results of the base classifiers as input values of the combining rule, so as to make better use of the information of sub-classifiers or sub-prediction models. In the experiments, six different triangular norms were applied separately for combination and prediction, they are Aczél–Alsina, Frank, Hamacher, Schweizer–Sklar, Sugeno–Weber and Yager t-norm. In the classification experiments, there are seven different datasets from the University of California, Irvine machine learning repository (UCI) were used for testing. Experimental results obtained on the datasets show that not all the triangular norm is suitable for combination. The performance of classification is improved effectively by employing the Yager t-norm operator, Schweizer-Sklar operator and Aczel-Alsina operator, and the Yager t-norm is best. In the prediction experiments, a dataset containing five data samples was used. For comparison, the BP prediction data and LS_SVM prediction data were used. Experimental results show that the new prediction rules based on t-norms can obtain better performance.

*Address correspondence to this author at the No. 58 Yanta Road, Xi'an, Shaanxi, 710054, China; Tel: +8613891939416;
E-mail: pengtao.jia@gmail.com

Table 1. Mathematical expressions of the used t-norms.

ID	Type	Binary Operator	Ternary Operator
1	Aczél–Alsina	$e^{-(\lvert\log x\rvert^p+\lvert\log y\rvert^p)^{1/p}}$	$e^{-(\lvert\log x\rvert^p+\lvert\log y\rvert^p+\lvert\log z\rvert^p)^{1/p}}$
2	Frank	$\log_p(1+\frac{(p^x-1)(p^y-1)}{p-1})$	$\log_p(1+\frac{(p^x-1)(p^y-1)(p^z-1)}{(p-1)^2})$
3	Hamacher	$\frac{xy}{p+(1-p)(x+y-xy)}$	$\frac{xyz}{p^2+p(1-p)(x+y+z)+(1-p)^2(xy+xz+yz)+(3p-p^2-2)xyz}$
4	Schweizer–Sklar	$\sqrt[p]{\max(0,x^p+y^p-1)}$	$\sqrt[p]{\max(0,x^p+y^p+z^p-1\ )}$
5	Sugeno–Weber	$\max(0,\frac{x+y-1+pxy}{1+p})$	$\max(0,\frac{(1+px)(1+py)(1+pz)-(1+p)^2}{p(1+p)^2})$
6	Yager	$\max(0,1-\sqrt[p]{(1-x)^p+(1-y)^p}\ )$	$\max(0,1-\sqrt[p]{(1-x)^p+(1-y)^p+(1-z)^p}\ )$

2. COMBINATION MODEL BASED ON T-NORMS

2.1. Introduction of Triangular Norms

In mathematics, a triangular norm (t-norm) is a kind of binary operation. It is mainly used in the framework of probabilistic metric spaces and in multi-valued logic, specifically in fuzzy logic.

Knowing from reference [8], a t-norm is a function T on the real unit interval [0, 1]. And it satisfies some important mathematical properties, as follows:

the associativity :

$T(x,T(y,z))=T(T(x,y),z)$

the boundary conditions:

$T(0,y)=0,\quad T(1,y)=y$

the commutativity:

$T(x,y)=T(y,x)$

the monotonicity:

$T(x1,y1)\le T(x2,y2)$

when $x1\le x2$ *and* $y1\le y2$

Using generators is a method for constructing t-norms. In this method, some known binary functions, such as addition and multiplication, can be transformed into a t-norm by using a unary function. The construction of t-norms by additive generators is based on the following theorem:

Theorem 1: Let $f:[0,1]\to[0,+\infty]$ be a strictly decreasing function, so for all x, y in [0,1], $f(1)=0$ and $f(x)+f(y)$ is in the range of f, equal to $f(0+)$ or +∞. Then the function $T:[0,1]\times[0,1]\to[0,1]$ is a t-norm when defined as $T(x,y)=f^{-1}(f(x)+f(y))$.

The function $f(x)$ in Theorem 1 is called the additive generators of $T(x,y)$. Different t-norms can be obtained by adopting different generators. Therefore, t-norm has good generalization ability.

Considering the complexity and objectivity of its application domain, t-norm needs some appropriate extensions, such as the parametric t-norms and the multivariate t-norms.

Parametric t-norms are proposed based on the fact that the *AND* and *OR* operations themselves can change continuously. They are extensions of the *AND* and *OR* operations, and they can be defined by explicit formulas depending on a parameter p.

According to the associativity:

$$t(x_1,x_2,\cdots,x_n)=t(t_{n-1}(x_1,x_2,\cdots,x_{n-1}),x_n)$$

multivariate t-norms can be extended frm the binary operators, so as to meet the needs of practical applications.

The mathematical expressions of the used t-norms are shown in Table **1**.

2.2. Construction of the Combination Model

Assume the base learners set is $T=\{C_1,C_2,\cdots,C_T\}$, the new combination model can be described as follows:

Regard the actual values as $y_0(t)\ \ (t=1,2,\cdots,n)$, n is the number of samples in dataset;

Regard the values obtained by $C_1,C_2,\cdots,C_T$ are $y_1(t),y_2(t),\cdots,y_T(t)$ respectively;

Regard the final values are $y(t)$. The combination model can be described as

Table 2. Brief descriptions of datasets.

NO.	Name	Attributes	Dataset Size	Classes	U_1	U_2	U_3
1	Haberman	3	306	2	154	76	76
2	Iris	4	150	3	75	39	36
3	Liver	7	345	2	173	86	86
4	Pima	8	768	2	384	192	192
5	Statlog heart	13	270	2	135	68	67
6	Wdbc	32	569	2	285	142	142
7	Wine	13	178	3	90	45	43

$$y(t) = F(\alpha_1 y_1(t), \alpha_2 y_2(t), \cdots, \alpha_T y_T(t)) \qquad (1)$$

Where, F is the combination rule, α_i is the weight of $y_i(t)$ and $\alpha_i \in [0,1]$ and $\sum_{i=1}^{T} \alpha_i = 1$.

Consider the t-norm operators in Table **1** as F. Taking the Schweizer–Sklar t-norm for instance, the combination model can be described as:

$$y(t) = \sqrt[p]{\alpha_1 y_1(t)^p + \alpha_2 y_2(t)^p + \cdots + \alpha_T y_T(t)^p - 1} \qquad (2)$$

Genetic algorithm (GA) is used to estimate the parameter p.

GA is a global optimization search algorithm that can do well in many optimization problems. It is formed by simulating the genetic and evolutionary principle of organisms and referencing the random statistical principle.

Given that the prediction errors are objective and inevitable, the errors between the actual values $y_0(t)$ and the predictive ones $y(t)$ should be investigated. Consider the error estimation as follows:

$$E = \sum_{t=1}^{n} (y(t) - y_0(t))^2 \qquad (3)$$

Minimizing E is used as the evaluation of the objective function in genetic algorithm, and then the parameter p in the combination model can be obtained.

3. EXPERIMENTS AND RESULTS

3.1. Ensemble Classification

In the classification experiments, the weights α_i were set as $\alpha_i = 1$, $i = 1, 2, \cdots, T$ and the parameter p is obtained by Genetic algorithm.

A. Datasets

To validate our model, the experiments are designed based on seven datasets, which are from the UCI repository of machine learning database [9]. Every dataset is divided into three subsets: U_1, U_2 and U_3 randomly, by using the GENDAT function. GENDAT function is mainly used to generate datasets for training and testing randomly in Pattern Recognition toolbox (PRtool) of MATLAB.

U_1 is used to train the single classifiers, and U_2 is used to test the single classifiers trained by U_1 and train the new ensemble rules, and U_3 is used to test the single classifiers, the new ensemble rules and other combination rules. Table **2** shows these datasets and their descriptions.

B. Experimental Procedure

The experimental procedure is listed as follows:

Step 1: There is discrepancy in the sequential values of the dataset. Thus the values of the datasets should be normalized in interval [0,1] for comparison. The formula of normalization is given as follows:

$$norm(x_i) = \frac{x_i - \min(X)}{\max(X) - \min(X)}$$

Step 2: The Binary Decision Tree classifier, k-Nearest Neighbor classifier and Naive Bayes classifier are chosen as component classifiers. They are named as BDT, KNN and NB shortly and k is set with three.

Step 3: U_1 is used to train the single classifiers, U_2 is used to test the single classifiers and to train the new ensemble rules based on t-norms.

Step 4: Genetic algorithm is used to evaluate parameters of t-norms in U_2.

Set the following parameters of GA for parameter optimization:

The initial population is 20;

Use binary coding with eight numbers;

Select operation by using the uniform distribution random model;

Table 3. Mathematical expressions of the used t-norms.

Name	Haberman	Iris	Liver	Pima	Statlog Heart	Wdbc	wine
BDT	0.266	0.091	0.427	0.342	0.375	0.069	0.144
NB	0.263	0.060	0.375	0.246	0.174	0.061	0.041
3-NN	0.283	0.040	0.360	0.287	0.353	0.076	0.047
Product	0.259	0.053	0.353	0.266	0.195	0.057	0.040
Mean	0.262	0.063	0.375	0.266	0.174	0.061	0.035
Median	0.265	0.076	0.381	0.320	0.342	0.066	0.054
Maximum	0.263	0.063	0.375	0.266	0.174	0.061	0.037
Minimum	0.265	0.090	0.381	0.320	0.342	0.066	0.080
Voting	0.256	0.050	0.345	0.271	0.250	0.050	0.040
Aczel-Alsina	0.247	0.033	0.144	0.219	0.191	0.030	0.060
Frank	0.255	0.056	0.362	0.290	0.296	0.058	0.102
Hamacher	0.329	0.047	0.367	0.367	0.296	0.059	0.088
Schweizer-Sklar	0.214	0.064	0.215	0.169	0.146	0.032	0.084
Sugeno-Weber	0.328	0.322	0.356	0.266	0.303	0.061	0.344
Yager	0.189	0.025	0.072	0.116	0.046	0.046	0.065

Do crossover operation by the disperse cross;

Mutate operation by using gauss function.

Step 5: U_3 is used to test the single classifiers, the new ensemble rules based on t-norms and other commonly used combination rules, such as product rule, mean rule, median rule, maximum rule, minimum rule [10], and majority voting rule.

Step 6: To compare all the results and analyze the effect of the new rules.

There are differences among the results of every experiment, because records of U_1, U_2 and U_3 are selected randomly. Repeat the experiment for ten times, and consider the mean-value of the ten experimental results as the final results for comparison.

C. Results Analysis and Discussion

In this section, the effect of six different t-norm operators on the classification results is investigated. It mainly focuses on the classification error rates, which are shown in Table **3**.

According to Table **3**, Aczel-Alsina operator and Schweizer-Sklar operator produces better performance on five datasets than the single classifiers and other combination rules, which are product rule, mean rule, median rule, max rule, min rule and voting rule. Yager operator produces better performance on six datasets. Specifically, five of them are the best among all the t-norms, and their results are 0.189, 0.046, 0.025, 0.072, 0.116 respectively.

So, it can be concluded that choosing the appropriate ensemble algorithm can improve the accuracy of classification. Among these t-norm operators, Yager operator is the most suitable operator used as the fusion rule of multiple classifiers.

3.2. Ensemble Prediction

In the prediction experiments, the weight α_i and the parameter p are all obtained by Genetic algorithm.

A. Datasets

The prediction experiments are designed on a dataset of coal spontaneous combustion from Datong, Xinzhou coal mine. As shown in Table **4**, the input data contains five influence factors of the output data, which is the ceiling intensity of air leakage in coal spontaneous combustion. And, the first 20 samples are regarded as the training set, the last five are regarded as the testing set.

B. Experimental Procedure

The experimental procedure is listed as follows:

Step 1: Normalize the data to interval [0,1] for comparison according to the formula in the classification experiments.

Step 2: Choose the BP neural network prediction model and the Least squares support vector machine forecasting model (LS_SVM) as the single prediction methods, and train them on the training set.

Table 4. Coal spontaneous combustion data used in the prediction experiments.

No.	Input Data					Output Data
	Distance/m	Oxygen Concentration/%	Coal Temperature /°C	Heat Intensity /10^5J·s^{-1}·cm^{-3}	Coal Thickness /m	Ceiling Intensity of Air Leakage/ cm^3·cm^{-2}·s^{-1}
1	1.70	20.60	19.60	0.87	7.0	0.70
2	2.50	20,04	20.30	1.04	6.0	0.83
3	4.70	19.88	22.00	1.27	5.0	1.08
4	7.60	19.03	22.50	1.34	4.0	1.56
5	16.30	18.21	24.20	1.43	2.0	2.35
6	20.50	17.99	25.60	1.51	3.0	2.58
7	25.20	17.60	26.70	1.58	4.0	2.88
8	29.10	17.36	26.80	1.58	3.0	3.17
9	36.40	16.90	27.50	1.62	2.0	3.43
10	43.90	15.74	28.30	1.67	3.0	3.87
11	44.30	15.68	28.60	1.69	4.0	3.92
12	47.00	14.91	28.10	1.66	5.0	4.12
13	53.70	13.77	25.13	1.49	7.0	5.60
14	56.40	13.09	24.80	1.47	6.0	5.77
15	59.00	12.44	24.30	1.43	5.0	6.00
16	61.20	11.93	23.60	1.40	4.0	6.53
17	70.60	10.78	24.67	1.46	2.0	5.39
18	74.30	9.81	26.30	1.55	2.0	4.18
19	78.00	8.85	27.80	1.61	3.0	3.76
20	89.20	7.14	30.40	1.79	3.0	2.89
21	11.00	18.59	23.40	1.39	3.0	1.88
22	39.70	16.50	27.90	1.64	2.0	3.52
23	50.40	14.36	28.20	1.66	6.0	4.24
24	66.80	11.18	24.20	1.43	3.0	6.01
25	83.50	7.97	28.10	1.66	4.0	3.76

Step 3: testing the BP and LS_SVM model on the testing set, and obtain their prediction data. Table **5** shows the actual testing values and the single prediction data.

Step 4: Regard the BP prediction data and LS_SVM prediction data as the input values of the new combination model.

Step 5: Genetic algorithm is used to evaluate the weights and parameter of the weighted prediction model. And set the same parameters of GA for parameter optimization as the classification experiments.

Step 6: Compare the prediction data obtained by the new model and the single prediction ones, then five error calculation methods are used to analysis the effect of the new rules.

The error calculation methods are: the Sum of Squares Error (SSE), the Mean Absolute Error (MAE), the Mean Square Error (MSE), the Mean Absolute Percentage Error (MAPE) and the Mean Square Percentage Error (MSPE).

Table 5. The actual testing values and the single prediction data.

Sample Number	Actual Value	BP Prediction Data	LS-SVM Prediction Data
1	1.88	1.81977	1.93764
2	3.52	3.66935	3.56403
3	4.24	4.12589	4.37216
4	6.01	6.01739	5.99676
5	3.76	3.79867	3.61776

Table 6. Error rates of the prediction experiments.

Name	SSE	MAE	MSE	MAPE	MSPE
BP prediction model	0.0405	0.0740	0.0403	0.0226	0.0121
LS_SVM prediction model	0.0431	0.0779	0.0415	0.0229	0.0118
Aczel-Alsina t-norm based rule	0.0005	0.0610	0.0082	0.0152	0.0043
Frank t-norm based rule	0.0121	0.0319	0.0220	0.0089	0.0062
Hamacher t-norm based rule	0.0013	0.0539	0.0000	0.0147	0.0033
Schweizer-Sklar t-norm based rule	0.0121	0.0316	0.0220	0.0089	0.0062
Sugeno-Weber t-norm based rule	0.0121	0.0319	0.0220	0.0089	0.0062
Yager t-norm based rule	0.0001	0.0626	0.0087	0.0160	0.0041

C. Results, Analysis and Discussion

In this section, the effect of six different t-norm operators on the prediction results is investigated. It mainly focuses on the prediction error rates, which are shown in Table **6**.

According to Table **6**, no matter which t-norm is used, the new fusion rule can get better forecast effect than the single predictions when the five error calculation methods are used.

Taking the SSE for example, BP and LS_SVM prediction models obtain the error rates 0.0405 and 0.0431 respectively. The new combination rules based on t-norms get the error rates 0.0005, 0.0121, 0.0013, 0.0121,0.0121 and 0.0001, which are lower than the single ones; in this case, the Yager t-norm based rule perform best. When the MAE is used, the Schweizer-Sklar t-norm based rule gets the lowest error rate 0.0316. When the MSE is used, the Hamacher t-norm based rule can predict the values without error. When the MAPE is used, the new rules based on Frank t-norm, Schweizer-Sklar t-norm and Sugeno-Weber t-norm obtain the same error rate 0.0089, which is lower than others. When the MSPE is used, the Hamacher t-norm based rule gets the lowest error rate 0.0033.

So, it can be concluded that the rules based on the parameterized t-norms can predict the values with higher accuracy, they are Suitable prediction.

CONCLUSION

As it is known, ensembles can lead to improved accuracy compared to a single classification or a single prediction. But only appropriate combination rule can make full use of the base classifiers and produce higher accuracy. This paper presents new combination rules based on t-norms, due to the better generalization ability of t-norms. The experimental results show that Aczel-Alsina t-norm operator, Schweizer-Sklar t-norm operator and Yager t-norm operator can improve the performance of ensemble learning. They are suitable for combining the results of base classifiers or prediction models and the Yager t-norm is the best one for classification.

CONFLICT OF INTEREST

The author confirms that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Project supported by the Scientific Research Program Funded by Shaanxi Provincial Education Department (Program No. 2013JK0870).

REFERENCES

[1] Z. H. Zhou, “Ensemble learning”, In: *Encyclopedia of Biometrics*, Li, S.Z. and A. Jain (Eds). Springer: Berlin, pp: 270-273, 2009.

[2] G. Tsoumakas, I. Partalas, and I. Vlahavas, “A taxonomy and short review of ensemble selection”, In: *Proceedings of the 18th European Conference on Artificial Intelligence*, Patras, Greece, 2008, pp. 1-6.

[3] R.O. Duba, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd Ed, Wiley: New York, 2001.

[4] A. H. R. Ko, R. Sabourin, and A. S. Britto Jr., “Form dynamic classifier to dynamic ensemble selection”, *Pattern Recog*., vol. 41, no. 5, pp. 1718-1731, 2008.

[5] Chitra, and S. Uma, “An ensemble model of multiple classifiers for time series prediction”, *Int. J. Comput. Theory Eng*., vol. 2, no. 3, pp. 1793-8201, June 2010.

[6] S. D. Bay, “Combining nearest neighbor classifiers through multiple feature subsets”, In: *Proceedings 15th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, pp. 37-45, 1998.

[7] F. Farahbod, and M. Eftekhari, “Comparison of different t-norm operators in classification problems”, *Int. J. Fuzzy Log. Syst*., vol. 2. no. 3, pp. 33-39, 2012.

[8] M. Mizumoto, “Pictorial representation of fuzzy connectives, part 1: cases of t-norms, t-conorms and averaging operators”, *Fuzzy Sets Syst*., vol. 31, pp. 217-242, 1989.

[9] J. Newman, S. Hettich, C. L. Blake, and C. J. Merz, UCI Repository of Machine Learning Databases. University of California, Department of Information and Computer Science, Irvine, CA. Available from: http://www.ics.uci.edu/~mlearn/ MLRepository.html, 1998.

[10] J. Kittler, M. Hatef, R. P. W. Duin and J. Matas, “On combining classifiers”, *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 20, no. 3, pp. 226-239, 1998.

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014