Research on the Controllable Confidence Regression

Fangchun Jiang

ShenZhen Institute of Information Technology, ShenZhen, GuangDong, 518172,China

email: 06112061@bjtu.edu.cn

Keywords: Machine Learning; Confidence Machine; Confidence Regression

Abstract. The confidence machine has been an integral part of the target of the research in the field of machine learning. Confidence regression is a significant research field of confidence machine . Performs error evaluation on results of regressive learning to classify the accept field and the refuse field so as to achieve the confidence regression. By setting specific error value, the Algorithm achieves controllable confidence regression, which has been tested on experimental data sets. Finally, the problems of current research are discussed, and the research direction is pointed out.

Introduction

Creditable machine learning result is always the pursuing and researching target of scientists. The so called creditable machine learning result means the machine learning algorithm not only gives judgment of 'YES'/'NO' or 'Belong to'/'Not belong to', but also the credibility and reliability of such judgment, therefore determine whether accept, modify or abort this judgment. This creditable machine learning method can be achieved through confidence mechanism within the machine learning algorithm. The machine learning algorithm with confidence mechanism is Confidence Machine or CM. The CM not only gives prediction but also gives qualified measurement to each sample's prediction, which is Confidence and Credibility [1]. The confidence and creditability could estimate the quality of prediction and determines process strategy according to it. Introducing confidence mechanism to construct confidence machine in machine learning is suitable and necessary for classification recognition in high risk occasions, such as medical diagnosis, large sophisticated system failure diagnosis and network intrusion detection [2][3][4].

Confidence mechanism is the strategy and method that simultaneously gives creditability of learning result at the time machine learning algorithm outputs learning result. Confidence mechanism determines the creditable degree and effect of confidence machine, the creditability that outputted by it is important proof when setting system threshold or deciding which countermeasures to take.

Learning method with refuse option confidence mechanism sets a threshold value at the time of classifying determination, and refuses those classes which may cause uncertainties and errors, minimizing the error rate, hence achieves confident determination.

First of all, divide the sampling space into two mutually complementary fields: refusal field R and acceptance field (or classifying field) A . Their definitions are as follows: $R=\{x \mid 1-\max_i p(\omega_i \mid x)>t\}$, $A=\{x \mid 1-\max_i p(\omega_i \mid x)\le t\}$, where t is the threshold value. The smaller the threshold t, the larger the refusal field R [5]. If sample x is located in acceptance field A, classify x according to some learning method. If sample x is located in refusal field R, refuse to classify x.

Reference [6] carried on research on classifier with refusal option, it proposed optimized classifier and refusal rule. To classifying problems, according to Bayesian learning method, if $p(\omega_1 \mid x)\ge(1-t)$, then x belongs to ω_1, if $p(\omega_2 \mid x)\ge(1-t)$, then x belongs to ω_2, otherwise refuse x. In previous inequations, t is the refusal threshold, and a constant $0\le t\le 0.5$, to set posterior probability threshold. If t's value is 0, all classified samples are refused. If t is 0.5, it is same as Bayesian learning rules without refusal option. The posterior probability that refuses all samples is always less than $1-t$. Certainly, this rule requires the probability distribution of known

samples, this is normally difficult or impractical in real world.

Reference [7] introduced support vector machine adopted "double hinge loss function" used in electrocardiogram classification. It used smallest classifying cost with refusal option, and achieved 99% sensitivity, which was higher than other separately published research result. Reference [8] proposed a new cascade random sub space fusion method with refusal option, which performed automatic mammary cancer diagnosis. The first stage used support vector classifier to classify the mammary gland image, then inputted the refused unrecognizable image to second stage classifier which had multi-layer sensor. The remaining unrecognizable images were left for human determination. Reference [9] also introduced some other researches on machine learning with refusal options to achieve learning result's creditability.

To sum up, the confidence mechanism with refusal option eliminates some low possibility instances using threshold value in order to lower the error rate, achieving creditable determination. Such creditability mechanism is suitable for those occasions need to analyze data with high cost, such as medical diagnosis and banknote verification. However, refused samples could be correct, and the follow-up processes of these samples still need further study.

Algorithm design

Support vector regression machine [10] includes two types, i.e., linear regression, and non-linear regression, whose basic idea is that data x are mapped to a high-dimensional characteristic space F through a non-linear mapping Φ, and linear regression is conducted in this space. Assuming that there are l training sample sets, i.e., $(x_1, y_1), (x_2, y_2), \cdots, (x_l, y_l) \in R^n \times R$, then the following can be obtained:

$$f(x) = (\omega \bullet \Phi(x)) + b$$

$$\Phi : R^n \to F, \quad \omega \in F \qquad \ldots(1)$$

where b is a threshold value. Through transformation in Eq. (1), the linear regression in the high-dimensional characteristic space corresponds to the non-linear regression in the low-dimensional input space, with the calculation of dot product of ω and $\Phi(x)$ in the high-dimensional space being omitted. Since Φ is fixed, ω is influenced by the sum of experience risks R_{emp} and the $\|\omega\|^2$ that flattens ω in the high-dimensional space, then the following can be obtained:

$$R(\omega) = R_{emp} + \lambda\|\omega\|^2 = \sum_{i=1}^{l} e(f(x_i) - y_i) + \lambda\|\omega\|^2 \qquad \ldots(2)$$

where l is the number of samples, $e(\bullet)$ is the loss function, and λ is a constant for adjustment. Through minimizing the $R(\omega)$, ω expressed with data points can be obtained:

$$\omega = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) \Phi(x_i) \qquad \ldots(3)$$

where α_i and α_i^* are the solutions of the minimized $R(\omega)$. Through Eqs. (1) and (3), $f(x)$ can be expressed as:

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)(\Phi(x_i) \bullet \Phi(x)) + b = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) k(x_i, x) + b \qquad \ldots(4)$$

where $k(x_i, x) = \Phi(x_i) \bullet \Phi(x)$ is called kernel function, and it is a dot product which meets the Mercer conditions in the characteristic space. There are many types of kernel functions, e.g., polynomial function $k(x_i, x) = [(x \bullet x_i) + 1]^q$, RBF function $k(x_i, x) = \exp\{-|x - x_i|^2 / 2\sigma^2\}$, and Sigmoid function $k(x_i, x) = \tanh(\nu(x \bullet x_i) + c)$.

The model used in the KNN algorithm corresponds to division of characteristic space in fact. The selection of k value may have relatively large influence on the results of the algorithm. If the k value is small, it indicates that only a training example that is relatively close to the input example can have an effect on the prediction results, thus over-fitting phenomenon easily occurs; if the k value is large, it can reduce the estimation error of learning, but the approximate error of learning increases, and training examples that are relatively far away from the input example can also have effects on the prediction, easily leading to wrong prediction.

The decision-making rule in the KNN algorithm is normally majority voting or all averaging, that is, the taken value of an input example depends on the k nearest training examples of the input example or is the average value of all examples.

Algorithm:

Input

X : The characteristic data value of sample set

Y : The regression value of sample set

Sample set (X, Y): $(\mathbf{x}_1,y_1), \ldots,(\mathbf{x}_l,y_l) \in R^n \times R$

Training Set Number: The number of samples in training set

Test Set Number: The number of samples in test set

k: The number of nearest neighbors taken

t: Error limit

Output

Test Set A: Test set acceptance region

Test Set R: Test set rejection region

MSE: Mean squared error before rejection

MSE-A: Mean squared error after rejection

Rejection rate (ReR): Rejection rate

Process

1. Conduct scale processing of data sets, and obtain optimized parameters c, g, and p, etc.;
2. Take samples randomly according to the Training Set Number and Test Set Number, using the sample set (X , Y) and through setting random function, to generate Training Set and Test Set;
3. Conduct training on the Training Set using the regression machine LIBSVM, to obtain a regression machine model;
4. Conduct regression prediction on the Test Set using the regression machine model;
5. Calculate the T value using the KNN algorithm;
6. According to the t value, determine the Test Set A, and Test Set R;
7. Calculate the MSE, and MSE-A;
8. Calculate the ReR;
9. Cycle the steps 2–8 for 10 times, and calculate the experimental data values of these 10 times;
10. End

Algorithm realization

For this algorithm, the tool LIBSVM [11] is used as a regression tool, and programming is conducted on the MATLAB7.0 to realize the algorithm. The data sets used in experiment are regression data sets in the UCI [12], etc., including data sets of bodyfat, housing, pyrim, triazines, and cpusmall, and the specific parameters of the data sets are shown in Table 1.

Table 1 Status of data sets used in experiment

No.	Data set name	Data set type	Number of characteristic value	Number of sample	Number of training set	Number of testing set
1	bodyfat	regression	14	252	200	52
2	housing	regression	13	506	400	106
3	pyrim	regression	27	74	50	24
4	triazines	regression	60	186	150	36
5	cpusmall	regression	12	8192	7000	1192

Normalization processing of data was conducted using the Scale method, the Scale range was [-1,1]; with bodyfat as an example, the command was as follows:

svm-scale -l -1 –u 1 -s range bodyfat > bodyfat.scale

And then, gridregression.py was used to identify optimal parameters; with bodyfat as an example, the command was as follows:

python gridregression.py -svmtrain "e:\libsvm-3.17\windows\svm-train.exe" -gnuplot "c:\program files\gnuplot\bin\gnuplot.exe" -log2c -10,10,1 -log2g -10,10,1 -log2p -10,10,1 -v 10 -s 3 -t 2 bodyfat.scale > gridregression_feature.parameter

The last line of the file gridregression_feature.parameter was outputted in the current path, and there were optimized parameters c, g, and p.

The parameters of the LIBSVM regression machine are shown in Table 2.

Table 2 Parameters of LIBSVM regression machine

No.	data sets	c	g	p	s	t
1	bodyfat	16.0	0.0009765625	0.0009765625	Type of svm,s=3 is e -SVR	Type of kernel function, t=2 is RBF kernel function
2	housing	256.0	0.25	2.0		
3	pyrim	512.0	0.0625	0.015625		
4	triazines	1.0	0.03125	0.0625		
5	cpusmall	256.0	0.5	6.0		

Finally, in the algorithm, the processed data sets were used to carry out experiment; the experiment was repeated for 10 times on each data set so as to obtain an average learning level, and the data obtained after the experiment were analyzed and compared.

Test results

Firstly, the experiment is conducted on the UCI provided bodyfat data set, which has 252 samples in total. The training set number is set as 200, and the test set number is set as 52. Table 3 shows the experimental results of LIBSVM on the bodyfat data set with the algorithm cycled for 10 times when t=0.01 and k was taken as 1–10, respectively.

Table 3 Experimental results of LIBSVM on bodyfat

k	t	ReR(%)	MSE	MSE-A
1	0.01	25	5.322046e-006	2.224215e-006
2	0.01	14.81	5.322046e-006	1.399697e-006
3	0.01	10.77	5.322046e-006	5.495417e-007
4	0.01	10.96	5.322046e-006	5.356577e-007
5	0.01	13.27	5.322046e-006	5.057531e-007
6	0.01	15.38	5.322046e-006	4.768819e-007
7	0.01	15.19	5.322046e-006	4.633491e-007
8	0.01	16.15	5.322046e-006	4.748111e-007
9	0.01	16.15	5.322046e-006	4.517478e-007
10	0.01	16.73	5.322046e-006	4.589683e-007

As can be seen from Table 3, in the case that the t value is fixed, when k is taken as 3, 4, and 5, the overall learning results differs little; when k is taken as 6–10, the overall learning levels are very close to each other. With the rejection rate and error being comprehensively taken into account, the result is ideal when k=3.

The data in Table 3 are mean values in 10 times of running, and all MSE-A values are less than MSE values, indicating that the algorithm is effective as a whole. In fact, for different k values, there are some running times out of 10 times of learning, when the MSE-A values of running results are larger than corresponding MSE values; there are 2 times for k=1, 3 times for k=2, and 1 time for each of k=3–10.

Conclusion

The confidence regression based on the KNN algorithm realizes controlled confidence regression, based on support vector regression machines, with the KNN algorithm as a tool, and setting different specific values of parameters t and k. It was verified on five experimental data sets including bodyfat data set, with good experimental effects being obtained.

Moreover, further research efforts on how to raise the control accuracy of controlled confidence regression and on how to more accurately select relevant parameter values including k and t are needed.

Acknowledgement

In this paper, the research was sponsored by the Natural Science Foundation of Guangdong (No. S2011010000824); Shen Zhen Collaborative Science and Technology Innovation Project (GJHS2012062809424003); Shenzhen Educational Scientific Projects (2014, gh004).

References

[1] Tony Bellotti, Zhiyuan Luo, Alex Gammerman. Reliable Classification of Childhood Acute Leukaemia from Gene Expression Data Using Confidence Machines[C]. Atlanta, GA: IEEE International Conference on Granular Computing, 2006. 148-153.

[2] Nouretdinov I, Costafreda SG, Gammerman A, Chervonenkis A, Vovk V, Vapnik V, Fu CHY. Machine learning classification with confidence: Application of transductive conformal predictors to MRI-based diagnostic and prognostic markers in depression[J]. Neuroimage, 2011, 56(2): 809-813.

[3] R Polikar, L UdPa, S UdPa, V Honavar. An incremental learning algorithm with confidence estimation for automated identification of NDE signals[J] . IEEE Transactions on Ultrasonics Ferroelectrics and Frequency Control, 2004, 51(8): 990-1001.

[4] LI Yang, et al. A novel network anomaly detection method based on TCM-KNN algorithm[J] . Journal of Software, 2007, 18(10): 2595-2604. (in Chinese)

[5] Andrew R Webb. Statistical Pattern Recognition (Second Edition) [M] . Beijing: Publishing House of Electronics Industry, 2004. (in Chinese)

[6] Chow C K. On optimum recognition error and reject tradeoff[J] . IEEE Trans. on Info. Theory, 1970, 16: 41-46.

[7] Zidelmal Z, Amirou A, Belouchrani A. Heartbeat classifcation using support vector machines (SVMs) with an embedded reject option[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2012, 26(1): 1250001-1-17.

[8] Zhang YG, Zhang BL, Coenen F, Lu WJ. Highly Reliable Breast Cancer Diagnosis with Cascaded Ensemble Classifiers[C]. Brisbane, Australla: WCCI 2012 IEEE World Congress on Computational Intelligence, 2012.

[9] Choi Hosik, Yeo Donghwa, Kwon Sunghoon, Kim Yongdai. Gene selection and prediction for cancer classification using support vector machines with a reject option[J] . Computational Statistics and Data Analysis, 2011, 55(5): 1897-1908.

[10] Vladimir N. Vapnik. Statistical Learning Theory [M]. Beijing: Publishing House of Electronics Industry, 2009: 406-413.

[11] http://www.csie.ntu.edu.tw/~cjlin/

[12] http://archive.ics.uci.edu/ml/