Distance Metric Learning-based Conformal Predictor

Fan Yang[1]，Zhigang Chen[1]，Guifang Shao[1]，Huazhen Wang[2]
[1]Department of Automation, Xiamen University, Xiamen, 361005, China
[2]School of Computer Science and Technology, Huaqiao University, Xiamen, 361021, China

Abstract. In order to improve the computational efficiency of conformal predictor, distance metric learning methods were used in the algorithm. The process of learning was divided into two stages: offline learning and online learning. Firstly, part of the training data was used in distance metric learning to get a space transformation matrix in the offline learning stage; Secondly, standard CP-KNN was conducted on the remaining training data with a nonconformity measure function defined by K nearest neighbors classifier in the transformed space. Experimental results on three UCI datasets demonstrate the efficiency of the new algorithm.

Keywords: conformal predictor; distance metric learning

1 Introduction

Most of the machine learning algorithms neglect or even ignore the reliability of the classifications and predictions. The conformal predictor (CP) algorithm focuses on the reliability and confidence of classification and provides a predefined confidence level for each prediction [1]. What's more, the CP algorithm is online by nature, so the CP algorithm has great superiority in practical applications [2-4].

In previous studies, K nearest neighbors (KNN) classifier was often used to design the nonconformity measure function for CPs, which is known as CP-KNN algorithm [6]. There are some disadvantages in CP-KNN [8]. Firstly, the nonconformity measure function of CP-KNN is designed in the original space, but the distance metric in the original space can't adapt to most of the practical problems. Secondly, the distance metric is difficult to determine, and Euclidean distance and other distance are not good for many classification problems. Thirdly, it needs to store large amounts of historical data, and frequently search and access the database, so it will cost a lot of time. Especially, when the data sets are huge, the calculation is intolerable in many practical applications.

There are some existing versions of CPs aiming to improve its computational efficiency, e.g. offline learning transductive confidence machine [5], which is conducted in an offline manner. However, in offline learning, the calibration of CPs can not be guaranteed theoretically. In [7,9], Wang HZ and Yang F et al. proposed a hybrid compression CP which used random forest to learn a proximity matrix on part of training data and apply CP-KNN in an online learning manner on the remaining data in the sample space defined by the random forest proximity matrix. Hence the compu-

This work was partially supported by the Fundamental Research Funds of China for the Central Universities under Grant No. 2010121065 and the Research Grant Council of Huaqiao University with Project No.09BS515.

tational efficiency was improved through the transfer of the online learning computational cost to the offline learning. Further, Yang F et al. propose to use adaptive kernel learning on part of the training data and also got both high predictive efficiency and computation efficiency on a fault detection dataset [8]. Different from [10], which use Multiple Kernel Learning (MKL) methodology to maximize efficiency in the CP framework, the above work mainly focused on improving the learning framework of CPs by dividing the learning process into two parts: offline learning and online learning. And this learning strategy can also be viewed as Mondrian CPs according to [1].

In this paper, we further present the distance metric learning based Conformal Predictor. Identically, the learning process was divided in to two stages. In the offline learning section, we used part of the training data sets to get a space transformation matrix by distance metric learning methods [11-13], and then design the nonconformity measure function with KNN and apply online CP-KNN in the new space.

2 Conformal Predictor

Conformal predictor (CP) is a transductive confidence machine (TCM) based on Kolmogorov randomness theory and conducted in online scheme. When one testing example is coming, we assume the example belongs to every possible class, and get several new test sequences. Randomness test is then applied to every test sequence. If the randomness level of a new sequence is relatively high, the corresponding class may be the true label of the testing example. CP treats all the classes whose randomness levels are greater than the specified risk level as prediction results. So CP is region prediction rather than point prediction of the traditional machine learning algorithms [1].The key problem of the CP algorithm is the design of nonconformity measures for examples, which is the degree of a testing example consistent with each class's distribution.

3 Distance Metric Learning Algorithms

The KNN rule works well in the design of nonconformity measure in CPs. However the performance of KNN depends crucially on the distance metric. The performances of the KNN could vary greatly on different distance metrics. Usually, Euclidean distances are used as a similarity measure, which is not always the case in most applications. And the distance metric should adapt to the particular applications. We expected that the nonconformity measure with KNN could be more effective in a space learned from the training set. In this paper, we used three distance metric learning methods to get the space transformation matrix. The methods are Large Margin Nearest Neighbors Classifier (LMNN) [11], Discriminative Component Analysis (DCA) [12] and Local Fisher Discriminant Analysis (LFDA) [13].

4 Distance Metric Learning-based Conformal Predictor Algorithm

In this paper, the process of learning is divided into two stages: offline learning and online learning. And the training data set is also randomly divided into two parts: compressed data set and calibration data set.

(i) Offline learning. In the original space, the compressed data set is used with Distance Metric Learning methods (LMNN, LFDA and DCA) to get a space transformation matrix (M). The purpose of this stage is to reduce storage size and the online computational cost by compressing part of the training data into knowledge.

(ii) Online learning. We get a new data space from offline learning stage. The transformation matrix not only increases the class separability of the compressed data set, but also increases the class separability of the calibration data set and the testing data set. The nonconformity measure function which is designed by KNN in the new space will be more effective.

The nonconformity measure function can be defined in the new space as:

$$\begin{aligned} \alpha_i &=: A_i([Z_1,...,Z_{i-1},Z_{i+1},...,Z_n],M), i=1,...,n-1 \\ \alpha_n &=: A_n([Z_1,...,Z_{n-1}],M) \end{aligned} \tag{1}$$

where Z_n is a testing example and M is the linear transformation matrix. A_i is the nonconformity measure function.

And the output of the prediction:

$$\Gamma^{\varepsilon}(Z_1,...,Z_{n-1},x_n) = \{y \in Y : p =| (i = m_k + 1,...,n : \alpha_i \geq \alpha_n) | /n > \varepsilon\} \tag{2}$$

The nonconformity measure score was calculated as:

$$\alpha_i = \sum_{j=1}^{k} \theta_{ij}^{-y} / \sum_{j=1}^{k} \theta_{ij}^{y}, i = m_k + 1,...,n, j \neq i \tag{3}$$

Where $\sum \theta_{ij}^{-y}$ denotes the proximity sum of the k nearest neighbors whose class label are not y, and $\sum \theta_{ij}^{y}$ denotes the proximity sum of the k nearest neighbors whose class label are y.

Algorithm 1. Distance Metric Learning-based Conformal Predictor

---

Input: $(x_1, y_1),...,(x_{n-1}, y_{n-1})$, testing data x_n , k, and the risk level ε ;

Output: the region prediction Γ^{ε} of x_n .

Step 1. The compressed data set is used with Distance Metric Learning methods to get a space transformation matrix (M);

Step 2. The calibration data set is put into the new space by the space transformation matrix (M);

Step 3. Initialize the region prediction set and the calibration data set:

$$\Gamma^{\varepsilon} = \Phi,$$

$$V = \{(x_{m+1}, y_{m+1}),...,(x_{n-1}, y_{n-1})\};$$

Step 4. For test example x_n, assume every class label $y \in \{1,2,...,c\}$ as its label, and calculate the nonconformity score of each assumption with Eq. (15). If the random level (p) of the nonconformity scores sequence is larger than the risk level (ε), $\Gamma^{\varepsilon} = \Gamma^{\varepsilon} \bigcup y$.

Step 5. Get the true class label (y_n) of x_n, and add x_n to the calibration data set $V = V \bigcup \{x_n, y_n\}$.

In our experiments we set k=3, ε = 80%,95%,99% respectively.

5 Experiment Results

We call the Distance Metric Learning-based Conformal Predictor Algorithms as CP-LMNN, CP-LFDA and CP-DCA respectively. We compared our new algorithms with CP-KNN on three UCI data sets, i.e., SPAM E-mail Database, the Pen-Based Recognition of Handwritten Digits and the Thyroid disease [14].

Four evaluation indicators are used, i.e., certain prediction rate, certain and correct prediction rate, empty prediction rate and exact calibration. The certain prediction rate means the rate of the region predictions which only have one class label. The certain and correct prediction rate denotes the rate of the region predictions which only have one label and the label is just the true class label. The empty prediction rate denotes the rate of the region predictions which have no elements.

Results on SPAM E-mail Database are showed in Fig.1. The results on the other two datasets are similar to Fig.1. And the certain and correct prediction rate are shown in Table 1.The experiments were performed on a laptop machine with a 2.1 GHz core 2 processor. In order to compare the computational efficiency, the CPU time for each experiment is listed in Table 2.

Table 1 shows that the certain and correct prediction rate of the new methods are better than CP-KNN at most confidence levels. The rate of region predictions which only have one label is improved, and most of the predictions are just the true class labels of the testing data. And the average degree of similarity of certain and correct prediction rate and certain prediction rate are also better.

From Table 2 we can see that the time costs of our method on all data sets are less than CP-KNN. Note that, the larger of the compressed data set, the CPU times of CP-LMNN will be less. The compressed data is used in offline learning and would not influence the computation efficiency of online learning. In contrast, as the number of historical data increases, CP-KNN will be very time-consuming. So the new algorithm is suitable for the online learning of huge data sets.

6 Conclusions

In this paper we propose a distance metric learning based conformal predictor. The new algorithm works well in improving the computational efficiency of CP. And at the same time, the predictive efficiency is also improved by some degree.

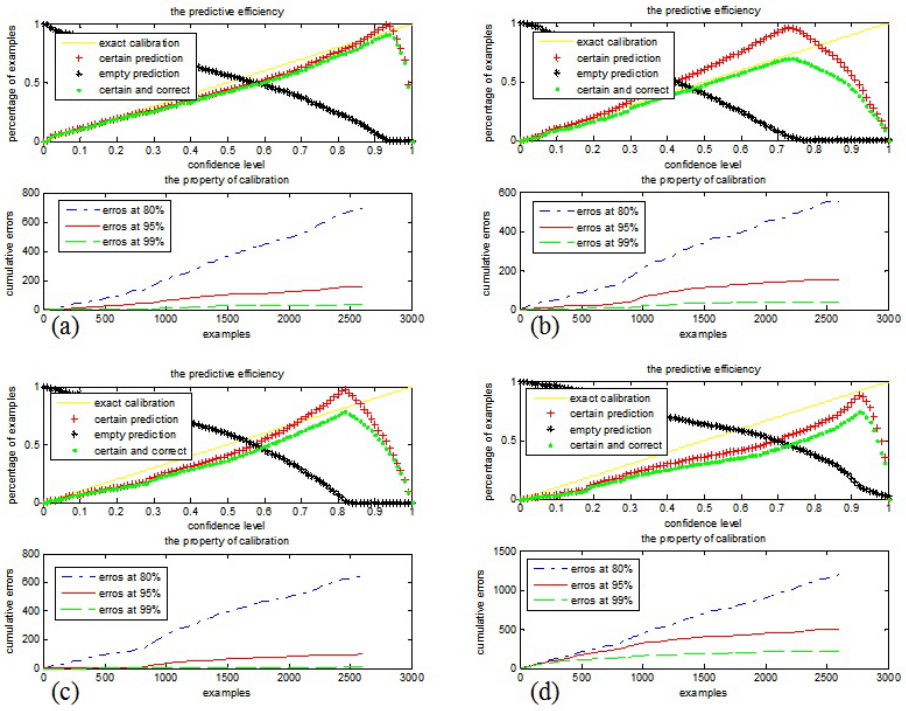

Fig. 1. The predictive efficiency on Spam E-mail Database, (a)CP-LMNN (b)CP-DCA (c)CP-LFDA (d)CP-KNN

Table 1. The Comparison of the predictive efficiency*

	CL(%)	0.1	0.3	0.5	0.7	0.8	0.9	Cc/c
Spam E-mail	KNN	3.6	19.49	30.6	43.37	53.86	69.86	86.58
	LMNN	10.77	25.99	42.71	60.44	73.32	87.04	96.46
	LFDA	6.81	20.53	36.68	61.74	75.12	58.25	89.66
	DCA	9.11	27.56	47.4	67.67	62.82	43.83	79.41
Hand-written	KNN	4.33	27	46.53	66.8	75.93	86.4	99.95
	LMNN	11.67	33.73	53.47	72.8	82.73	91	99.86
	LFDA)	10.07	31.87	55.47	75.47	84.2	92.2	99.32
	DCA	10.8	31.13	52.2	71.8	82.2	92.33	99.41
Thyroid	KNN	11.5	28.33	50.09	69.11	79.1	87.64	92.67
	LMNN	10.52	29.59	52.55	74.98	84.2	81.65	95.14
	LFDA)	10.52	30.17	54.07	75.33	84.15	73.68	93.45
	DCA	9.71	30.48	53.22	75.16	84.42	72.11	93.02

*CL: confidence level; Cc/c: certain and correct prediction rate/certain prediction rate

Table 2. The Comparison on the Time Cost*(s)

Dataset	N	CM	CA	TE	Attribute	LMNN	LFDA	DCA	KNN
E-mail	4601	1000	1000	2601	58	4697	4139	3987	7409
Hand-written	4500	2000	1000	1500	17	5849	6257	6437	15778
Thyroid	4534	1150	1150	2234	22	9126	8833	9442	16873

*N : the number of examples used in the experiments; CM (compressed): the number of training data used for distance metric learning; CA (calibration): number of remaining training data used for online learning. TE :the test data.

References

1. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic learning in a random world. Springer, New York (2005).
2. Gammerman, A., Vovk, V.: Prediction algorithms and confidence measures based on algorithmic randomness theory. Theoretical Computer Science. 287,209-217(2002).
3. Gammerman, A., Vovk, V.: Hedging predictions in machine learning. Computer Journal. 50,151-177(2007).
4. Vovk, V.: A Universal Well-Calibrated Algorithm for On-line Classification. J. Mach. Learn. Res. 5,575-604(2004).
5. Stijn, V., Laurens, V.D.M., Ida S.K.: Off-line learning with transductive confidence machines: an empirical evaluation. In: Petra Perner (eds.), LNAI,vol.4571,pp.310-323,Leipzig, Germany,Springer(2007)
6. Papadopoulos, H., Vovk, V., Gammerman, A.: Regression Conformal Prediction with Nearest Neighbours. J. Artif. Intell. Res. 40, 815-840(2011)
7. Wang H.Z., Lin C.D., Yang F., Zhuang J.F. An online Algorithm with confidence for Real-Time Fault Detection. Journal of Information and Computational Science, 6(1),305-313 (2009)
8. Yang F., Luo J., Wang H.Z., Peng Y.Q., Mi H. Optimized Kernel-Based Conformal Predictor for Online Fault Detection. Journal of Tianjin University, 42(7), 614-621(2009)
9. Yang, F., Wang，H.Z., Mi, H., Lin, C.D., Cai, W.W. Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. BMC Bioinformatics, 10, S1(2009)
10. Balasubramanian, V., Ye, J., Chakraborty, S., Panchanathan, S. Kernel Learning for Efficiency Maximization in the Conformal Predictions Framework. In: 9th International Conference on Machine Learning and Applications, Washington DC(2010)
11. Kilian, Q., Weinberger, Lawrence, K., Saul. Distance Metric Learning for Large Margin Nearest Neighbor Classification. J. Mach. Learn. Res.10,207-244(2009)
12. Peltonen, J., Goldberger, J., Kaski, S. Fast discriminative component analysis for comparing examples. Learning to Compare Examples-NIPS 2006 Workshop, 12(2006)
13. Masashi, S. Dimensionality Reduction of Multimodal Labeled Data by Local Fisher Discriminant Analysis. J. Mach. Learn. Res. 8, 1027-1061(2007)
14. Frank, A., Asuncion, A. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.