33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Confidence-Based Classification with Dynamic Conformal Prediction and Its Applications in Biomedicine

Yurong Luo, Abed Al-Raoof Bsoul, and Kayvan Najarian, Senior Member, IEEE

***Abstract*—Computer-aided decision support systems enable physicians to make more accurate clinical decisions and can significantly improve the quality of care provided to patients. However, prediction of classification confidence as the degree of reliability on the resulting predictions is a much needed step in clinical decision making. A recently developed technique called conformal prediction utilizes the similarity between a new sample and the training samples in order to form confidence measures for predictions. However, the conventional conformal prediction method suffers from shortcomings such as high computational complexity that prevent its use in real-time applications. This paper introduces an alternative approach to the conventional confidence prediction that addresses some of this and other disadvantages. Both real clinical and non-clinical datasets are employed to test and validate the capabilities of the proposed approach.**

I. Introduction

Over the last twenty years, computer-aided decision support systems [1, 2] have been developed as a result of increasing demand for management and processing of medical data. However, the majority of such systems only output a set of predictions without indicating their reliability, while in clinical applications the level of confidence over these predictions are of paramount importance.

A small group of classification techniques have introduced some type of confidence measure over their predictions. One such approach is the theory of Probably Approximately Correct (PAC) learning [3, 4] that produces upper bounds on the probability of error with respect to some confidence level. But these bounds are usually weak, over-pessimistic, and not practically useful. Another method is the Bayesian approach [5] to obtain strong confidence bounds. However, the Bayesian methods require some a priori assumptions about the distribution generating the data. Some other approaches for estimating confidence-related measure are also used, including calculation of the rate of errors, specificity, sensitivity etc.

Conformal prediction (CP) [6] is a recently developed technique for the same purpose. Specifically, using a base classifier, CP assigns to a new sample a label that would group the sample with the most similar examples previously seen, and then uses the degree of similarity to and within the previously seen examples to estimate the prediction confidence. Several definitions of similarity have been introduced to form confidence [7, 8, 9], and many different base classifiers have been used, including as Decision Trees, Neural Network [10, 11], and Support Vector Machine [7, 8, 9, 12, 13]. However, CP suffers greatly from computational inefficiency. Although Inductive Conformal Prediction (ICP) [10, 14], can partially relieve this disadvantage of CP, it does not address the root cause of the problem.

It has to be mentioned that the basic idea of CP is different from other modern machine learning algorithms, such as Neural Network, Decision Tree, and Support Vector Machine (SVM) as CP produces multiple decision rules/outputs in order to form the final prediction, while modern machine learning algorithms only predict one output. This paper aims at bridging between CP and modern machine learning algorithms while providing a new efficient CP confidence measures for prediction. To do this, an alternative structure, Dynamic Conformal Prediction (DCP), is proposed in the paper. While many classifier can be used in DCP, in this paper SVM is adopted as base classifier, and SVM-based DCP is formulated in a computationally efficient approach. Both non-clinical and clinical datasets are employed to verify the efficiency and accuracy of the proposed approach. Experimental results show that the proposed method outperform the CP in accuracy, time complexity, and confidence estimation.

The rest of this paper is organized as follows. In Section II, the main ideas of CP are briefly described. CP using SVM is described in Section III. Section IV details the proposed alternative approach. Finally, Results and Conclusion will be presented in Section V and VI.

II. Conformal Prediction

A. Introduction

Conformal prediction treats classification problem as follows. As in any machine learning algorithms, a set of training samples is given as follows:

$$(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l) \quad y_i \in \{-1,1\} \quad x_i \in R^n, \tag{1}$$

where x_i's are the input patterns, y_i's are the binary output classes, n is the dimension of the input space, and l is the number of training examples. Then, a set of testing samples are given as:

$$x_{l+1}, \dots, x_{l+k}. \tag{2}$$

where k is the number of testing samples. The aim of the learning task is to form predictions $y_{l+1}, \dots, y_{l+k}$ for the testing samples. CP assumes that there is a sample space Z, which is composed of elements $z \in Z$, where z is a sequence of input-output pairs, i.e.

$$(x_1, y_1), \dots, (x_l, y_l), (x_{l+1}, y^*_{l+1}), \dots, (x_{l+k}, y^*_{l+k}) \tag{3}$$

Note that z contains the training classified examples and the testing samples with their provisional classifications. CP also

Manuscript received March 26, 2011. This material is based upon work supported by the National Science Foundation under Grant No. IIS0758410.

Yurong Luo is with the Virginia Commonwealth University, Richmond, VA 23220 USA (phone: 804-615-7229; e-mail: luoy4@vcu.edu).

Abed Al-Raoof Bsoul was with the Virginia Commonwealth University, Richmond, VA 23220 USA. He is now with ECPI College of Technology (e-mail: bsoulaak@vcu.edu).

Kayvan Najarian is with Virginia Commonwealth University, Richmond, VA 23220 USA (e-mail: knajarian@vcu.edu).

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

assumes that both training and testing datasets are generated independently from the same distribution Q. The labels of testing samples are predicted so as to make the samples most similar to the training samples according to a specific criterion which is a computable function on Z. The same function is employed as the basis to form "confidence".

B. Problems

Assuming k $(k \ge 1)$ testing samples and r labels for each sample, the number of all possible labels for the testing samples will be r^k when all k testing samples are presented to the learning algorithm at the same time. This means that in theory the procedure needs be run r^k times to find the most similar prediction and form the final classification. If only one testing sample is considered at each time step, the procedures will be run $r \times k$ times. Therefore, if k is too large a number, the learning process becomes computational expensive. In this paper, it is assumed that only one testing sample is presented to the learning algorithm at each time.

III. SUPPORT VECTOR MACHINE-BASED CONFORMAL PREDICTION

A. Support vector machine

Due to its popularity as a machine learning algorithm, support vector machine (SVM) [15] is adopted as the base classifier in this paper. The objective of SVM is to use a linear separating hyper plane to create a classifier with maximal margin. Considering training data are given as

$$(x_1, y_1),(x_2, y_2)\ldots(x_l, y_l), x \in R^n, y \in \{-1,+1\} \quad (4)$$

Formally, this is done by the minimum of objective function *E*,

$$E = \frac{1}{2}w^T w + C\left(\sum_{i=1}^{l}\xi_i\right) \quad (5)$$

subject to the following inequality constraints:

$$y_i\left[w^T x_i + b\right] \ge 1 - \xi_i, i = 1, l, \xi_i \ge 0 \quad (6)$$

Here *C* is penalty factor, *w* are weights, *b* is the bias, and ξ_i are non-negative slack variables.

The SVM maps the input vectors $x \in R^n$ into vectors z of a higher-dimensional feature space F in order to solve a non-linear classification problem in the original feature space. The mapping is realized using the kernel functions:

$$K(x_i, x_j) = \phi^T(x_i)\phi(x_j) \quad (7)$$

The solution in a dual space becomes the solution of the following problem (α_i is the Lagrange multiplier):

$$L(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}y_i y_j \alpha_i \alpha_j K(x_i, x_j) \quad (8)$$

$$C \ge \alpha_i \ge 0 \quad \text{and} \quad \sum_{i=1}^{l}\alpha_i y_i = 0 \quad (9)$$

Finally, SVM finds final hyper planes in the following form:

$$d(x) = \sum_{i=1}^{l}y_i \alpha_i K(x, x_i) + b \quad (10)$$

B. Conformal prediction based on Support vector machine

Consider the training samples / testing sample pairs in which the label of testing sample is provisional, i.e.

$$(x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, y_{l+1}^*)$$

The distribution Q generating both training and testing samples is usually considered to be exchangeable [6]. Each example in this set has an associated conformity score or non-conformity score. Both conformity and non-conformity scores can be interpreted as a measure of 'supportiveness' of the example. A high non-conformity score as well as a low conformity score indicate that the example is "strange" and unlikely to occur. Based on conformity and non-conformity scores, the p-value is constructed during classification stage as the measure of how strange the test example. More specifically, the p-value identifies the probability of observing the particular ordering of the conformity score, ∂^{con}, or non-conformity score, $\partial^{non-con}$, under the assumption that the testing sample is correctly classified [8]. The rank of ∂ is often used to form p-value due to the fact that if the rank of ∂_{new} is n, it means that ∂_{new} is the n *th* highest or lowest value. Specifically, using non-conformity, the p-value is defined as:

$$p - value = P\left\{rank(\partial^{non-con}{}_{new}) \ge n\right\} \quad (11)$$

While if conformity is used, the p-value is defined as

$$p - value = P\left\{rank(\partial^{con}{}_{new}) \le n\right\} \quad (12)$$

which are equivalent respectively to

$$p - value = \frac{\#\{i : \alpha_i \ge \alpha_{new}\}}{l+1} \quad (13)$$

and

$$p - value = \frac{\#\{i : \alpha_i \le \alpha_{new}\}}{l+1} \quad (14)$$

If the sample has r labels, after running the learning algorithm r times, there will be r p-values, and CP will assign the testing sample the provisional label with the highest p-values.

CP defines two criteria to estimate the prediction: credibility and confidence [7, 8, 9]. Credibility indicates the quality of the data on which the decision is based, while confidence tells the confidence in prediction. Credibility is defined as the largest p-value, while confidence is defined as

$$confidence = 1 - p_2 \quad (15)$$

where p_2 is the second largest p-value.

To better understand the concept of confidence, let us tentatively choose a 'significance level' such as 1%. If the confidence in prediction exceeds 99% but the actual prediction is wrong, the sample sequence belongs to a priori chosen set of data with probability less than 1%. As summarized in [12], in a reliable prediction: 1) confidence is relatively high, and 2) credibility is not low.

Conformal prediction based on SVM can use two measures, Lagrange multiplier α (non-conformity) and distance d (conformity). Regarding α, there are three possible solutions for α.

1. $\alpha = 0$: Data point is correctly classified.

2. $C > \alpha > 0$: Data point is called free support vector, i.e. it lies on the two margins (for linear classifier of non-overlapping classes, $C = \infty$).

3. $\alpha = C$: Data point is called bounded support vector, i.e. it lies on the wrong side of the margin.

Parameter d refers to the distance from data point to hyper plane. The distance becomes larger as the data point moves further from the hyper plane.

IV. DYNAMIC CONFORMAL PREDICTION

A. Main concept

Dynamic conformal prediction (DCP), designed in this paper, provides not only higher accuracy and lower computational complexity but also more confidence in the resulting predictions. The structure of DCP is presented in Fig.1. In the structure, the set of training samples is iteratively updated after a pre-specified time. That is, the system continues to bring in new training samples, and at the same time desert some “older” training samples, as shown in the dashed box. This is essential in many time-varying systems such as biomedical applications where the nature of the system/data changes as the time evolves. Therefore, the criteria to 'abandon' and 'add' is time. After processing via base classifier, the prediction gives not only the label of the testing samples but also the confidences in prediction, as in CP. However, another difference between DCP and the conventional CP in confidence prediction is four folds: 1) DCP just use credibility as useful information. The reason is that, credibility and confidence vary in a scope to indicate the reliability of prediction because each time the decision rule will change. So, it is relatively difficulty to set threshold to alarm the prediction. In fact, credibility can give the same hints about prediction. 2) A new confidence measure is designed and used in DCP, 3) DCP uses only distance d as conformity score, 4) Unlike the conventional CP, the proposed DCP deals with multi testing samples using one decision rule in order to form the final prediction which makes DCP more inline with other machine learning.

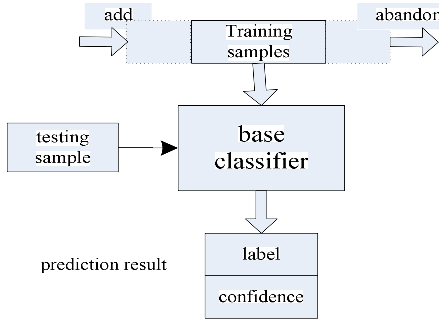


Fig.1 Dynamic Conformal Prediction

In CP, based on each testing sample, the conformity or the non-conformity score obtained from base classifier is used to predict the label and confidence. As mentioned above, this approach in conformal prediction is computationally inefficient and expensive. On the other hand, when the decision rule/output is formed using the majority of machine learning methods, conformity scores (d) have been already acquired. As such, the DCP’s calculates confidences for multi samples, and investigates if it can achieve higher accuracy and the same level of confidence. As it will be shown in the experimental results, the computational complexity of the DCP is significantly less that CP.

B. Confidence

In this paper, a new form of confidence is also designed to show how reliable each of predictions is. The new form of confidence is defined as follows:

$$\text{confidence} = \arctan(\mathit{conformity}) \qquad (16)$$

The introduced definition, while reflecting the same level of information as the original one, has the following advantages: 1) The range of [0, $\pi/2$], normalized to [0,1] in our analyses, makes it much easier to set thresholds. In addition, each confidence level clearly reflects the relation between the data point and the hyper plane, 2) It is much less affected by the distribution of data (as shown in formula (13) and (14)).

V. RESULT

A. Datasets

Both clinical and non-clinical datasets are employed to test and verify the effectiveness of our proposed method. Non-clinical datasets (five datasets 'australian' (690/14), 'breast-cancer' (683/10), 'diabetes' (768/8), 'heart' (270/13), 'liver-disorders' (345/6), 'splice' (1000/60)) all from Lib-SVM repository (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/) were used. Clinical datasets (94088 samples with 17 features) are based on MIT-BIT [16] to detect arrhythmia.

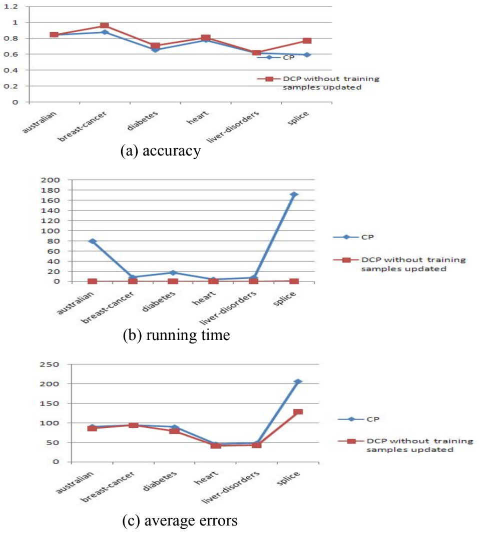
B. Experimental Result

For base classifier, SVM, the best parameters are set, which are selected based on ten-fold cross validation.

1) Non-clinical applications:

The nature of data has no time-sequence, thus DCP without updated training samples was used. The results (Fig.2) of accuracy, time complexity, and average errors are presented in (a), (b) and (c), respectively. The purpose of average errors (errors: the number of samples if correctly classified if confidences are below its current confidence, and wrongly classified if confidences are above its current confidence.) is to test how reliable the confidence measure is. The results show that the accuracies, running time and average errors using proposed method are superior to the original method.

In Fig. 3 horizontal axis stands for conformity or non-conformity score; and the vertical axis shows confidence. Fig. 3 shows the confidences curves of all data for both the original method (Fig 3.a and 3.c) and our proposed method (Fig 3.b and 3.d). Five datasets ('australian', 'breast-cancer', 'diabetes', 'heart', 'liver-disorders') in six have relatively similar results for both methods because the confidence increases as conformity (or non-conformity) score increases. However, one dataset ('splice') shows abnormal behavior in the original method (Fig 3.c) as when conformity (or non-conformity) score is around the same value, confidence does not show a clear pattern. However, our measure for the same dataset (Fig. 3.d) shows a clear and well-accentuated pattern of confidence.

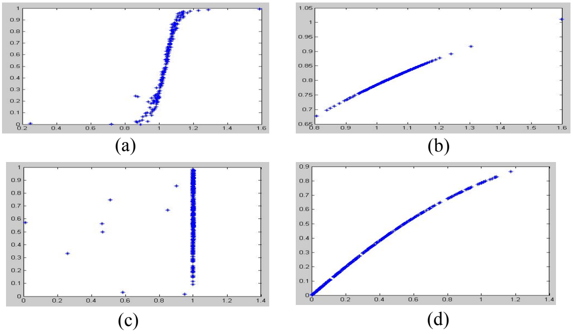


(a) accuracy

(b) running time

(c) average errors

Fig.2 The results of accuracy, time consumption and average errors



(a)

(b)

(c)

(d)

Fig.3 The shape of confidences for all data in one dataset

2) Clinical application

The total number of ECG beats in the MIT-BIT dataset is 94088 beats, which include 54088 training samples and 40000 testing samples. For testing samples, 28987 instances are healthy beats and 11013 instances are arrhythmia beats. First, feature extraction is performed on each beat using the methodology described in [17], and totally 17 features are extracted. Then using the extracted features, DCP with updated training samples is used for classification. The resulting accuracy is 98.23% which is significantly higher than that of the original CP which is 94.75%. Also, the time complexity in our proposed method is significantly reduced from 2750 seconds (in CP) to 5.4063 seconds on the same machine with the same hardware.

VI. Conclusion

In order to assist physicians in make more accurate decision, a measure of confidence on prediction is desired in any computer-aided decision support system. Based on the similarity between a new sample and the previously observed samples, CP produces a prediction and a confidence measure. However, CP's time complexity and lack of adaptation makes it unsuitable for many real applications. In order to overcome these shortcomings, an alternative structure, DCP, is proposed. The proposed DCP provides the multiple advantages over CP as it deals with multi-testing samples as in other machine learning methods. Moreover, a new form of confidence is proposed that unlike CP, is based on the main idea of conformity score. In addition, the new form of confidence is not influenced by the distribution of data points, and is confined to the interval of [0, 1], which makes it much easier for threshold setting. In the future, we intend to form new measures for conformity and non-conformity scores and use them to further address the issue of time complexity.

References

[1] Sholom M. Weiss, Casimir A. Kulikowski, Saul Amarel and Aran Safir, "A Model-Based Method for Computer-Aided Medical Decision Making", Artificial Intelligence, pp.145-172, 1978.

[2] S.Y. Ji, R. Smith, T. Huynh and K. Najarian, "A Comparative Analysis of Multi-level Computer-Assisted Decision Making Systems for Traumatic Injuries", *BMC Medical Informatics and Decision Making* 9:2, 2009.

[3] David Haussler, "Probably Approximately Correct Learning", AAAI-90 Proceedings, pp.1101-1108, 1990.

[4] David Haussler, "Overview of the Probably Approximately Correct Learning Framework"

[5] Thomas Melluish, Craig Saumders, Ilia Nouretdinov, and Volodya Vovk, "Comparing the Bayes and typicalness frameworks", Proceedings 12th European Conference on Machine Learning (ECML'01), Lecture Notes in Computer Science, vol.2167, pp.360-371. Springer (2001)

[6] Vladimir Vovk, Alexander Gammerman, Glenn Shafer, "Algorithmic learning in a Random World", Springer, New York, 2005

[7] A.Gammerman, V.Vovk, V.Vapnik, "Learning by Transduction", In Uncertainty in Artificial Intelligence, pp.148-155, 1998

[8] Craig Saunders, Alex Gammerman, Volodya Vovk, "Transduction with confidence and credibility", In Proceedings of the International Joint Conference on Artificial Intelligence, pp.722-726, 1999.

[9] Volodya Vovk, Alex Gammerman, Craig Saunders, "Machine-Learning Applications of Algorithmic Randomness", In Proceedings of ICML'99, pp.444-453, 1999

[10] Harris Papadopoulos, Alex Gammerman, Colodya Vovk, "Confidence Predictions for the Diagnosis of Acute Abdominal Pain", IFIP Advances in information and Communication Technology, pp.175-184, 2009

[11] Harris Papadopoulos, Volodya Vovk, Alex Gammerman, "Conformal Prediction with Neural Networks", 19th IEEE International Conference on Tools with Artificial Intelligence, pp.388-395, 2007

[12] Alex Gammerman, Volodya Vovk, "Prediction algorithms and confidence measures based on algorithmic randomness theory", Theoretical Computer Science, pp.209-217, 2002

[13] Qiu De-hong, Chen Chuan-bo, JIN Xian-ji, "Confidence Support Vector Machine Based on Algorithmic Theory of Randomness and its Application on Signature Verification", MINI-MICRO SYSTEMS, pp.2131-2134, 2004

[14] Harris Papadopoulos, Vladimir Vovk, Alex Gammerman, "Qualified predictions for large data sets in the case of pattern recognition", Proceedings 2002 International Conference on Machine Learning and Applications, pp.159-163, 2002.

[15] Vojislav Kecman. "Learning and Soft Computing: support vector machine, neural networks, and fuzzy logic models". The MIT Press, Cambridge, MA, pp.120-184, 2001.

[16] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/cgi/content/full/101/23/e215]; 2000 (June 13).

[17] A.A.R. Bsoul, S.Y. Ji, K. Ward, and K. Najarian, Automatic Prediction of Arrhythmia Severity Using Time and Frequency Domain Features, Circulation 122:A213, 23 November 2010