Confidence Predictions for the Diagnosis of Acute Abdominal Pain

Harris Papadopoulos, Alex Gammerman and Volodya Vovk

Abstract Most current machine learning systems for medical decision support do not produce any indication of how reliable each of their predictions is. However, an indication of this kind is highly desirable especially in the medical field. This paper deals with this problem by applying a recently developed technique for assigning confidence measures to predictions, called *conformal prediction*, to the problem of acute abdominal pain diagnosis. The data used consist of a large number of hospital records of patients who suffered acute abdominal pain. Each record is described by 33 symptoms and is assigned to one of nine diagnostic groups. The proposed method is based on Neural Networks and for each patient it can produce either the most likely diagnosis together with an associated confidence measure, or the set of all possible diagnoses needed to satisfy a given level of confidence.

1 Introduction

Machine learning techniques have been applied successfully to many medical decision support problems [7, 8] and many good results have been achieved. The resulting systems learn to predict the diagnosis of a new patient based on past history of patients with known diagnoses. Most such systems produce as their prediction only the most likely diagnosis of the new patient, without giving any confidence

Harris Papadopoulos
Computer Science and Engineering Department, Frederick University, 7 Y. Frederickou St., Palouriotisa, Nicosia 1036, Cyprus. e-mail: H.Papadopoulos@frederick.ac.cy

Alex Gammerman
Department of Computer Science, Royal Holloway, University of London, Egham Hill, Egham, Surrey TW20 0EX, England. e-mail: Alex@cs.rhul.ac.uk

Volodya Vovk
Department of Computer Science, Royal Holloway, University of London, Egham Hill, Egham, Surrey TW20 0EX, England. e-mail: Vovk@cs.rhul.ac.uk

information in this prediction. This is a major disadvantage, as measures of confidence are of paramount importance in a medical setting [6]. Confidence measures are an indication of how likely each prediction is of being correct. In the ideal case, a confidence of 99% or higher for all examples in a set, means that the percentage of erroneous predictions in that set will not exceed 1%; when this is true we say that the confidence measures are well calibrated.

Conformal prediction (CP) [24] is a recently developed technique, which can be used for obtaining confidence measures. Conformal predictors are built on top of traditional machine learning algorithms, called *underlying algorithms*, and complement the predictions of these algorithms with measures of confidence. Different variants of CPs are described in [11, 15, 16, 17, 18, 21, 22, 23]. The results reported in these papers show that not only the confidence values output by CPs are useful in practice, but also their accuracy is comparable to, and sometimes even better than, that of traditional machine learning algorithms.

Of course other approaches that can be used for deriving some kind of confidence information do exist. One can apply the theory of Probably Approximately Correct learning (PAC theory) to an algorithm in order to obtain upper bounds on the probability of its error with respect to some confidence level. These bounds though, are usually very week [12] and as a result not very useful in practice. Another alternative is the use of Bayesian methods which can give strong confidence bounds. Bayesian methods however, require some a priori assumptions about the distribution generating the data and if these are violated their outputs can become quite misleading [10].

In this paper we apply CP to the problem of acute abdominal pain diagnosis. This is a relatively popular problem in medical decision support, see e.g. [2, 3, 5, 9, 13, 20, 25], due to the poor discrimination between the diseases that cause acute abdominal pain, which results in high diagnostic error rates [25]. Wrong diagnoses may result in unnecessary emergency abdominal operations, or in complications, such as perforation of the appendix.

The CP we use is based on Neural Networks (NNs). NNs have not only been successfully applied to many medical problems [1, 4, 8, 19], but they are also one of the most popular machine learning techniques for almost any type of application. In order to use NNs as the underlying algorithm of a CP, we follow a modified version of the original CP approach called Inductive Conformal Prediction (ICP) [14]. ICP is based on the same general idea as CP but, as its name suggests, it replaces the transductive inference used in the original approach with inductive inference. ICP was first proposed in [16, 17] in an effort to overcome the computational inefficiency problem of CPs. As demonstrated in [18] this computational inefficiency problem renders the original CP approach highly unsuitable for use with NNs; and in extend any other method that requires long training times.

The rest of this paper is structured as follows. In section 2 we summarise the general idea behind CP and its inductive version ICP, while in section 3 we detail the Neural Network ICP method. Section 4 gives an analysis of the data used in this study and section 5 describes our experiments and lists and discusses their results. Finally, section 6 gives our conclusions and the future directions of this work.

2 Conformal Prediction

In this section we give a brief description of the idea behind CP, for more details see [24]. We are given a training set $\{z_1,\ldots,z_l\}$ of examples, where each $z_i \in Z$ is a pair (x_i,y_i); $x_i \in \mathbb{R}^d$ is the vector of attributes for example i and $y_i \in \{Y_1,\ldots,Y_c\}$ is the classification of that example. We are also given a new unclassified example x_{l+1} and our task is to state something about our confidence in each possible classification of x_{l+1}.

CP is based on measuring how likely it is for each extended set of examples

$$\{(x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},Y_j)\} : j = 1,\ldots,c, \tag{1}$$

to have been generated independently from the same probability distribution. First we measure how strange, or non-conforming, each example in (1) is for the rest of the examples in the same set. We use what is called a *non-conformity measure* which is based on a traditional machine learning algorithm, called the *underlying algorithm* of the CP. This measure assigns a numerical score α_i to each example (x_i,y_i) indicating how different it is from all other examples in (1). In effect we train the underlying algorithm using (1) as training set and we measure the degree of disagreement between its prediction for x_i and the actual label y_i; in the case of x_{l+1} we use the assumed label Y_j in the place of y_{l+1}.

The non-conformity score $\alpha_{l+1}^{(Y_j)}$ of (x_{l+1},Y_j) on its own does not really give us any information, it is just a numeric value. However, we can find out how unusual (x_{l+1},Y_j) is according to our non-conformity measure by comparing $\alpha_{l+1}^{(Y_j)}$ with all other non-conformity scores. This comparison can be performed with the function

$$p((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},Y_j)) = \frac{\#\{i = 1,\ldots,l+1 : \alpha_i \geq \alpha_{l+1}^{(Y_j)}\}}{l+1}. \tag{2}$$

We call the output of this function, which lies between $\frac{1}{l+1}$ and 1, the p-value of Y_j, also denoted as $p(Y_j)$, as that is the only part of (1) we were not given. An important property of (2) is that $\forall\delta \in [0,1]$ and for all probability distributions P on Z,

$$P^{l+1}\{((x_1,y_1),\ldots,(x_{l+1},y_{l+1})) : p(y_{l+1}) \leq \delta\} \leq \delta; \tag{3}$$

for a proof see [12]. As a result, if the p-value of a given label is under some very low threshold, say 0.05, this would mean that this label is highly unlikely as such sequences will only be generated at most 5% of the time by any i.i.d. process.

After calculating the p-value of every possible label Y_j, as described above, we are able to exclude all labels that have a p-value under some very low threshold (or *significance level*) δ and have at most δ chance of being wrong. Consequently, given a confidence level $1-\delta$ a CP outputs the set

$$\{Y_j : p(Y_j) > \delta\}. \tag{4}$$

Alternatively the CP can predict the most likely classification together with a confidence and a credibility measure in this prediction. In this case it predicts the classification with the largest p-value, outputs one minus the second largest p-value as confidence to this prediction and as credibility it outputs the p-value of the predicted classification, i.e. the largest p-value.

2.1 Inductive Conformal Prediction

The original CP technique requires training the underlying algorithm once for each possible classification of every new test example. This means that if our problem has 9 possible classifications and we have to classify 2000 test examples, as is the case in this study, the training process will be repeated $9 \times 2000 = 18000$ times. This makes it very computationally inefficient especially for algorithms that require long training times such as Neural Networks.

Inductive Conformal Predictors (ICPs) are based on the same general idea described above, but follow a different approach which allows them to train their underlying algorithm just once. This is achieved by splitting the training set (of size l) into two smaller sets, the *proper training set* with $m < l$ examples and the *calibration set* with $q := l - m$ examples. The proper training set is used for training the underlying algorithm and only the examples in the calibration set are used for calculating the p-value of each possible classification of the new test example. More specifically, we calculate the p-value of each possible classification Y_j of x_{l+1} as

$$p(Y_j) = \frac{\#\{i = m+1, \ldots, m+q, l+1 : \alpha_i \geq \alpha_{l+1}^{(Y_j)}\}}{q+1}, \tag{5}$$

where $\alpha_{m+1}, \ldots, \alpha_{m+q}$ are the non-conformity scores of the examples in the calibration set and $\alpha_{l+1}^{(Y_j)}$ is the non-conformity score of (x_{l+1}, Y_j).

3 Neural Networks Inductive Conformal Predictor

In this section we analyse the Neural Networks ICP (NN-ICP) algorithm. We first describe the typical output encoding for Neural Networks (NNs) and then, based on this description, we define two non-conformity measures for NNs. Finally, we detail the complete NN-ICP algorithm.

3.1 Non-conformity Measures

Typically the output layer of a classification NN consists of c units, each representing one of the c possible classifications of the problem at hand; thus each label is encoded into c target outputs. To explicitly describe this encoding consider the label, $y_i = Y_u$ of a training example i, where $Y_u \in \{Y_1, \ldots, Y_c\}$. The resulting target outputs for y_i will be

$$t_1^i, \ldots, t_c^i,$$

where

$$t_j^i = \begin{cases} 1, \text{ if } j = u, \\ 0, \text{ otherwise,} \end{cases}$$

for $j = 1, 2, \ldots, c$. In the same manner we will denote the actual outputs of the NN for an example i as

$$o_1^i, \ldots, o_c^i.$$

According to this encoding the higher the output o_u^i (which corresponds to the example's true classification) the more conforming the example, and the higher the other outputs the less conforming the example. In fact, the most important of all other outputs is the one with the maximum value $\max_{j=1,\ldots,c: j \neq u} o_j^i$, since that is the one which might be very near or even higher than o_u^i. So a natural non-conformity measure for an example $z_i = (x_i, y_i)$ where $y_i = Y_u$ would be defined as

$$\alpha_i = \max_{j=1,\ldots,c: j \neq u} o_j^i - o_u^i, \tag{6}$$

or as

$$\alpha_i = \frac{\max_{j=1,\ldots,c: j \neq u} o_j^i}{o_u^i + \gamma}, \tag{7}$$

where the parameter $\gamma \geq 0$ in the second definition enables us to adjust the sensitivity of our measure to small changes of o_u^i depending on the data in question. We added this parameter in order to gain control over which category of outputs will be more important in determining the resulting non-conformity scores; by increasing γ one reduces the importance of o_u^i and consequently increases the importance of all other outputs.

3.2 The Algorithm

We can now use the non-conformity measure (6) or (7) to compute the non-conformity score of each example in the calibration set and each test set pair (x_{l+g}, Y_u). These can then be fed into the p-value function (5), giving us the p-value for each classification Y_u. The exact steps the Neural Networks ICP follows for a training set $\{z_1, \ldots, z_l\}$ and a test set $\{x_{l+1}, \ldots, x_{l+r}\}$ are:

- Split the training set into the *proper training set* with $m < l$ examples and the *calibration set* with $q := l - m$ examples.
- Use the proper training set to train the Neural Network.
- For each example $z_{m+t} = (x_{m+t}, y_{m+t})$, $t = 1, \ldots, q$ in the calibration set:
  - supply the input pattern x_{m+t} to the trained network to obtain the output values $o_1^{m+t}, \ldots, o_c^{m+t}$ and
  - calculate the non-conformity score α_{m+t} of the pair (x_{m+t}, y_{m+t}) by applying (6) or (7) to these values.
- For each test pattern x_{l+g}, $g = 1, \ldots, r$:
  - supply the input pattern x_{l+g} to the trained network to obtain the output values $o_1^{l+g}, \ldots, o_c^{l+g}$,
  - consider each possible classification Y_u, $u = 1, \ldots, c$ and:
    - compute the non-conformity score $\alpha_{l+g} = \alpha_{l+g}^{(Y_u)}$ of the pair (x_{l+g}, Y_u) by applying (6) or (7) to the outputs of the network,
    - calculate the p-value $p(Y_u)$ of the pair (x_{l+g}, Y_u) by applying (5) to the non-conformity scores of the calibration examples and $\alpha_{l+g}^{(Y_u)}$:

$$p(Y_u) = \frac{\#\{i = m+1, \ldots, m+q, l+g : \alpha_i \geq \alpha_{l+g}^{(Y_u)}\}}{q+1},$$

    - predict the classification with the largest p-value (in case of a tie choose the one with the smallest non-conformity score) and output one minus the second largest p-value as confidence to this prediction and the p-value of the output classification as its credibility,
    - or given a confidence level $1 - \delta$ output the prediction set (4).

4 Acute Abdominal Pain Data

The acute abdominal pain database used in this study was originally used in [5], where a more detailed description of the data can be found. The data consist of 6387 records of patients who were admitted to hospital suffering from acute abdominal pain. During the examination of each patient 33 symptoms were recorded, each of which had a number of different discrete values. For example, one of the symptoms is "Progress of Pain" which has the possible values: "Getting Better", "No Change", "Getting Worse". In total there are 135 values describing the 33 symptoms. These values compose the attribute vector for each patient in the form of 135 binary attributes that indicate the absence (0) or presence (1) of the corresponding value. It is worth to mention that there are symptoms which have more than one value or no value at all in many of the records.

There are nine diseases or diagnostic groups in which the patients were allocated according to all information after their initial examination, including the results of

Table 1 Data distribution.

	APP	DIV	PPU	NAP	CHO	INO	PAN	RCO	DYS	Total
Training Set	585	108	88	1941	372	290	65	326	612	4387
Test Set	259	35	42	894	200	127	31	147	265	2000
Total	844	143	130	2835	572	417	96	473	877	6387

surgical operations. These are: Appendicitis (APP), Diverticulitis (DIV), Perforated Peptic Ulcer (PPU), Non-specific Abdominal Pain (NAP), Cholesistitis (CHO), Intestinal Obstruction (INO), Pancreatitis (PAN), Renal Colic (RCO) and Dyspepsia (DYS). NAP is not actually a diagnostic group, it is a residual group in which all patients that did not belong to one of the other groups were placed.

The data are divided into a training set consisting of 4387 examples and a test set consisting of 2000 examples. These are the same training and test sets as in [5]. Table 1 reports the number of examples that belong to each diagnostic group.

5 Experiments and Results

The NN used in our experiments was a 2-layer fully connected feed-forward network, with sigmoid hidden units and softmax output units. It consisted of 135 input, 35 hidden and 9 output units. The number of hidden units was selected by following a cross validation scheme on the training set and trying out the values: 20, 25, 30, 35, 40, 45, 50, 55, 60. More specifically, the training set was split into five parts of almost equal size and five sets of experiments were performed, each time using one of these parts for evaluating the NNs trained on the examples in the other four parts. For each of the five test parts, a further 10-fold cross validation process was performed to divide the examples into training and validation sets, so as to use the validation examples for determining when to stop the training process. Training was performed with the backpropagation algorithm minimizing a cross-entropy loss function.

The results reported here were obtained by following a 10-fold cross validation procedure on the training set in order to divide it into training and validation examples. To create the calibration set of the ICP, 299 examples were removed from the training set before generating the 10 splits. This experiment was repeated 10 times with random permutations of the training examples. Here we report the mean values of all 100 runs.

Table 2 reports the accuracy of the NN-ICP and original NN methods and compares them to that of the Simple Bayes, Proper Bayes and CART methods as reported in [5]. Additionally it compares them to the accuracy of the preliminary diagnoses of the hospital physicians, also reported in [5]. Both the original NN and NN-ICP outperform the other three methods and are almost as accurate as the hos-

Table 2 Predictive Accuracy of NN-ICP Compared to Other Methods.

Method	Correct Diagnoses (%)
Neural Networks ICP	75.74
Original Neural Networks	75.87
Simple Bayes	74
Proper Bayes	65
Classification Tree (CART)	65
Physicians (preliminary diagnoses)	76

pital physicians. As was expected the original NN performs slightly better than the ICP due to the removal of the calibration examples from the training set, however the difference between the two is negligible. This is a very small price to pay considering the advantage of obtaining a confidence measure for each prediction.

Table 3 lists the results of the NN-ICP when producing set predictions for the 99%, 95%, 90% and 80% confidence levels. More specifically it reports the percentage of examples for which the set output by the ICP consisted of only one label, of more than one label or was empty. It also reports in the last column the percentage of errors made by the ICP, i.e. the percentage of sets that did not include the true classification of the example. The values reported here reflect the difficulty in discriminating between the 9 diseases. Nevertheless, the set predictions output by the NN-ICP can be very useful in practice since they pinpoint the cases where more attention must be given and the diagnostic groups that should be considered for each one. Bearing in mind the difficulty of the task and the 76% accuracy of the preliminary diagnoses of physicians, achieving a 95% of accuracy by considering more than one possible diagnosis for only about half the patients is arguably a good result.

Table 3 NN-ICP Set Prediction Results.

Non-conformity Measure	Confidence Level	Only one label (%)	More than one label (%)	No label (%)	Errors (%)
(6)	99%	23.76	76.24	0.00	0.95
	95%	46.62	53.38	0.00	4.10
	90%	62.38	37.62	0.00	8.59
	80%	82.22	17.78	0.00	16.94
(7)	99%	25.80	74.20	0.00	0.95
	95%	47.58	52.42	0.00	3.75
	90%	65.32	34.68	0.00	8.11
	80%	87.32	12.38	0.30	17.23

6 Conclusions and Future Work

We have presented the application of a recently developed technique, called Conformal Prediction, to the problem of acute abdominal pain diagnosis. Unlike most conventional algorithms, our approach produces confidence measures in its predictions which are provably valid under the general i.i.d. assumption. Our experiments demonstrate that the Neural Networks ICP is very successful at this very difficult task, since its predictions are almost as accurate as the preliminary diagnoses of hospital physicians and its confidence measures are well calibrated and practically useful. The set predictions produced by NN-ICP identify the cases that require more attention as well as the most likely diagnoses of these cases.

One undesirable aspect of the data used in this study is the huge difference in the number of examples that belong to each class. For this reason, in the future we plan to repeat our experiments with an artificially balanced version of the training set created by performing random resampling of the training examples. Additionally, our directions for future research include further experimentation with other datasets for acute abdominal pain and with more non-conformity measures based on other popular algorithms such as support vector machines, decision trees and evolutionary techniques.

Acknowledgements This work was supported by the Cyprus Research Promotion Foundation through research contract PLHRO/0506/22 ("Development of New Conformal Prediction Methods with Applications in Medical Diagnosis").

References

1. Anagnostou, T., Remzi, M., Djavan, B.: Artificial neural networks for decision-making in urologic oncology. Review in Urology **5**(1), 15–21 (2003)
2. Anastassopoulos, G.C., Iliadis, L.S.: Ann for prognosis of abdominal pain in childhood: Use of fuzzy modelling for convergence estimation. In: Proceedings 1st International Workshop on Combinations of Intelligent Methods and Applications, pp. 1–5 (2008)
3. Blazadonakis, M., Moustakis, V., Charissis, G.: Deep assessment of machine learning techniques using patient treatment in acute abdominal pain in children. Artificial Intelligence in Medicine **8**(6), 527–542 (1996)
4. Christoyianni, I., Koutras, A., Dermatas, E., Kokkinakis, G.: Computer aided diagnosis of breast cancer in digitized mammograms. Computerized Medical Imaging and Graphics **26**(5), 309–319 (2002)
5. Gammerman, A., Thatcher, A.: Bayesian diagnostic probabilities without assuming independence of symptoms. Methods of Information in Medicine **30**(1), 15–22 (1991)
6. Holst, H., Ohlsson, M., Peterson, C., Edenbrandt, L.: Intelligent computer reporting 'lack of experience': a confidence measure for decision support systems. Clinical Physiology **18**(2), 139–147 (1998)
7. Kononenko, I.: Machine learning for medical diagnosis: History, state of the art and perspective. Artificial Intelligence in Medicine **23**(1), 89–109 (2001)
8. Lisboa, P.: A review of evidence of health benefit from artificial neural networks in medical intervention. Neural Networks **15**(1), 11–39 (2002)

9. Mantzaris, D., Anastassopoulos, G., Adamopoulos, A., Gardikis, S.: A non-symbolic implementation of abdominal pain estimation in childhood. Information Sciences **178**(20), 3860–3866 (2008)
10. Melluish, T., Saunders, C., Nouretdinov, I., Vovk, V.: Comparing the Bayes and Typicalness frameworks. In: Proceedings 12th European Conference on Machine Learning (ECML'01), *Lecture Notes in Computer Science*, vol. 2167, pp. 360–371. Springer (2001)
11. Nouretdinov, I., Melluish, T., Vovk, V.: Ridge regression confidence machine. In: Proceedings 18th International Conference on Machine Learning (ICML'01), pp. 385–392. Morgan Kaufmann, San Francisco, CA (2001)
12. Nouretdinov, I., Vovk, V., Vyugin, M.V., Gammerman, A.: Pattern recognition and density estimation under the general i.i.d. assumption. In: Proceedings 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory, *Lecture Notes in Computer Science*, vol. 2111, pp. 337–353. Springer (2001)
13. Ohmann, C., Moustakis, V., Yang, Q., Lang, K.: Evaluation of automatic knowledge acquisition techniques in the diagnosis of acute abdominal pain. Artificial Intelligence in Medicine **8**(1), 23–36 (1996)
14. Papadopoulos, H.: Tools in Artificial Intelligence, chap. 18. Inductive Conformal Prediction: Theory and Application to Neural Networks, pp. 315–330. I-Tech, Vienna, Austria (2008). URL http://intechweb.org/downloadpdf.php?id=5294
15. Papadopoulos, H., Gammerman, A., Vovk, V.: Normalized nonconformity measures for regression conformal prediction. In: Proceedings IASTED International Conference on Artificial Intelligence and Applications (AIA 2008), pp. 64–69. ACTA Press (2008)
16. Papadopoulos, H., Proedrou, K., Vovk, V., Gammerman, A.: Inductive confidence machines for regression. In: Proceedings 13th European Conference on Machine Learning (ECML'02), *Lecture Notes in Computer Science*, vol. 2430, pp. 345–356. Springer (2002)
17. Papadopoulos, H., Vovk, V., Gammerman, A.: Qualified predictions for large data sets in the case of pattern recognition. In: Proceedings 2002 International Conference on Machine Learning and Applications (ICMLA'02), pp. 159–163. CSREA Press (2002)
18. Papadopoulos, H., Vovk, V., Gammerman, A.: Conformal prediction with neural networks. In: Proceedings 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'07), vol. 2, pp. 388–395. IEEE Computer Society (2007)
19. Pattichis, C., Christodoulou, C., Kyriacou, E., Pattichis, M.: Artificial neural networks in medical imaging systems. In: Proceedings 1st MEDINF International Conference on Medical Informatics and Engineering, pp. 83–91 (2003)
20. Pesonen, E., Eskelinen, M., Juhola, M.: Comparison of different neural network algorithms in the diagnosis of acute appendicitis. International Journal of Bio-Medical Computing **40**(3), 227–233 (1996)
21. Proedrou, K., Nouretdinov, I., Vovk, V., Gammerman, A.: Transductive confidence machines for pattern recognition. In: Proceedings of the 13th European Conference on Machine Learning (ECML'02), *Lecture Notes in Computer Science*, vol. 2430, pp. 381–390. Springer (2002)
22. Saunders, C., Gammerman, A., Vovk, V.: Transduction with confidence and credibility. In: Proceedings of the 16th International Joint Conference on Artificial Intelligence, vol. 2, pp. 722–726. Morgan Kaufmann, Los Altos, CA (1999)
23. Saunders, C., Gammerman, A., Vovk, V.: Computationally efficient transductive machines. In: Proceedings of the Eleventh International Conference on Algorithmic Learning Theory (ALT'00), *Lecture Notes in Artificial Intelligence*, vol. 1968, pp. 325–333. Springer, Berlin (2000)
24. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, New York (2005)
25. Zorman, M., Eich, H.P., Kokol, P., Ohmann, C.: Comparison of three databases with a decision tree approach in the medical field of acute appendicitis. Studies in Health Technology and Informatics **84**(2), 1414–1418 (2001)