Feature Selection by Conformal Predictor

Meng Yang, Ilia Nouretdinov, Zhiyuan Luo, Alex Gammerman

Feature Selection by Conformal Predictor

Meng Yang, Ilia Nouretdunov, Zhiyuan Luo, Alex Garmmerman

Computer Learning Research Centre, Royal Holloway, University of London
Egham Hill, Egham, Surrey TW20 0EX, UK

Abstract. In this work we consider the problem of feature selection in the context of conformal prediction. Unlike many conventional machine learning methods, conformal prediction allows to supply individual predictions with valid measure of confidence. The main idea is to use confidence measures as an indicator of usefulness of different features: we check how many features are enough to reach desirable average level of confidence. The method has been applied to abdominal pain data set. The results are discussed.
Keywords: feature selection, conformal predictor, confidence estimation.

1 Introduction

When we deal with classification or regression problems, the size of the training data and noise in the data may affect the speed and accuracy of the learning system. Are all the features really important or can we use less features to achieve the same or better results? The irrelevant features will induce greater computational cost and may lead to overfitting. For example, in the domain of medical diagnosis, our purpose is to infer the relationship between the symptoms and their corresponding diagnosis. If by mistake we include the patient ID number as one input feature, an over-tuned machine learning process may come to the conclusion that the illness is determined by the ID number.

Feature selection is the process of selecting a subset of features from a given space of features with the intention of meeting one or more of the following goals.

1. Choose the feature subset that maximises the performance of the learning algorithm.
2. Minimise the size of the feature subset without reducing the performance of a learning problem significantly.
3. Reduce the requirement for storage and computational time to classify data.

The feature selection problem has been studied by the statistics and machine learning communities for many years. Many methods have been developed for feature selection and these methods can basically be classified into three groups: filter, wrappers and embedded feature selection [1, 2]. The filter method employs a feature ranking function to choose the best features. The ranking function gives a relevance score based on a sequence of examples. Intuitively, the more relevant the feature, the higher its ranking. Either a fixed number of (at most) t features with the highest ranking are selected, or a variable number of features above a

preset threshold are selected. Filter methods have been successful in a number of problem domains and are very efficient. Wrapper methods are general-purpose algorithms that searches the space of feature subsets, testing performance of each subset using a learning algorithm. The feature subset that gives the best performance is selected for final use. Some learning algorithms include an embedded feature selection method. Selecting features is then an implicit part of the learning process. This is the case, for example, with decision tree learners like ID3 [3] that use an information measure to choose the best features to make a decision about the class label. Other learning algorithms have been developed with embedded feature selection in mind. For example, Littlestone's WINNOW algorithm is an adaptation of the Perceptron algorithm [4] that uses multiplicative weight updates instead of additive.

The methods we mentioned above often use accuracy as criterion to select features; in our paper, we consider confidence for the feature selection. This is because in our approach we can regulate the accuracy by choosing a certain confidence level. We will use conformal predictors as a tool to perform feature selection. Conformal predictors are recently developed machine learning algorithms which supply individual predictions with valid measure of confidence [5]. Level of confidence in predictions produced by such algorithm can used as a performance measure instead of just accuracy [1].

Conformal predictors could be proceed in on-line and batch mode. In batch mode, a fixed size of training set will be used, we may get good results by chance. In on-line mode, the size of training set grows after prediction, it could consider all different sizes of the training set. So we will extend on-line conformal predictors in order to get the average confidence, make feature selection and present results of application of this approach on a medical database.

2 Conformal Predictor

Conformal predictor is a method that not just makes predictions but also provides corresponding confidences [5]. When we use this method, we predict labels for new objects and use the degree of conformity to estimate the confidence in the predictions.

We start by defining the concept of a nonconformity measure which is a way of measuring how well an example fits to a set. A measure of fitness is introduced by a nonconformity measure A. For a sequence $z_1, z_2, ...z_n$ of examples, where the i_{th} example z_i is composed of objects and labels, $z_i = (x_i, y_i)$, x_i means objects and y_i means label, $x_i \subset X$ and $y_i \subset Y$. We can write $\wr z_1, z_2, ..., z_n \wr$ for the bag consisting of the examples $z_1, z_2, ..., z_n$, we can score a distance between z_i and the bag $\wr z_1, z_2, ..., z_n \wr / z_i$, expressed by $\alpha_i = A(\wr z_1, z_2, ..., z_n \wr / z_i, z_i)$, called the nonconformity score.

Non-conformity score can be based on a classical algorithm of prediction, in this paper we use the nearest neighbors algorithm [6]. The idea of using the nearest neighbors algorithm to measure the nonconformity of example z, (x, y), from the other examples is comparing x's distance to other examples' objects

with the same label to its distance to others with different label.

$$\alpha = \frac{distance\ to\ z's\ nearest\ neighbor\ to\ other\ examples\ with\ the\ same\ label}{distance\ to\ z's\ nearest\ neighbor\ to\ other\ examples\ with\ a\ different\ label}$$

2.1 Prediction and Confidence

Assume that an i.i.d (independent and identically distributed) data are given: $z_1, z_2, ...z_{n-1}$. Now, we have an new example x_n and want to predict its label y_n. First of all, we give y_n a value which belongs to Y, then calculate the nonconformity score for each example, finally, compare α_n to the other α by using p-value: $\frac{|\{j=1,...n:\alpha_j \geq \alpha_n\}|}{n}$. If the p-value is small, then z_n is nonconforming, if it is large, then z_n is very conforming. After we tried every value in Y, each of the possible labels will get a p-value $p(y)$, then the label with the largest p-value will be our prediction and its corresponding confidence will be equal to 1− the second largest p-value.

2.2 Validity

In conformal predictor, we also could set a level of significance ϵ to find the prediction region which contains all labels $y \subset Y$ with the corresponding p-value larger than ϵ, that means we will have confidence $1 - \epsilon$ in our prediction about y, and the probability of the event that true label is not contained by prediction region should be ϵ. And we define the situation when the true label is not contained in prediction region as error. Thus, the prediction in on-line mode is under the guarantee and valid.

$$Prob\{pvalue \leq \epsilon\} \leq \epsilon$$

According to this, size of prediction region could be another criterion, for a specific significant level, smaller region sizes provide us more efficient predictions. In this paper, we will use number of uncertain prediction to express this property, and uncertain prediction means the situation when prediction region contains multi-classes.

The confidence corresponds to the minimal level, at which we can guarantee that the prediction is certain (consists of the only label) under i.i.d. assumption. For example, if $Y = \{1, 2, 3\}$, and $p(1) = 0.23$, $p(2) = 0.07$, $p(3) = 0.01$ is the results for one of the examples in on-line mode, the prediction is: 1 with confidence 0.93. We can say that true label is 1 or 2 if we wish to be right with probability 0.95, on this level we are not sure that it is 1. On the other hand, if it enough for us to be right with probability 0.90, then we can claim that it is 1. See [5] for details of conformal (confident) prediction.

We can find that if the prediction is very conforming, its confidence will be very high, the prediction is made depends on the objects of examples, which means the more useful features we use, the higher confidence we will get. So, we could use confidence to justify how useful the features are.

3 Data Description

We use abdominal pain dataset, it has 9 kinds of diagnosis as labels and 33 types of symptoms as object [7, 10], which are sex, age, pain-site onset, pain-site present, aggravating factors, relieving factors, progress of pain, duration of pain, type of pain, severity, nausea, vomiting, anorexia, indigestion, jaundice, bowel habit, micturition, previous pain, previous surgery, drugs, mood, calor, abdominal movements, abdominal scar, abdominal distension, site of tenderness, rebound, guarding, rigidity, abdominal masses, murphy's test, bowel sounds and rectal examination. Each of symptoms contains different numbers of values, we can give two options for each value, 1 and 0, which means the patient has this value of one symptom or not, separately. After this step, we get 135 features in total. There are around 6000 examples in the original dataset where some of them have missing values, so we use 1153 of them which do not have missing values.

List of diagnostic groups is below:

Diagnostic Groups		
Group	Diagnosis	Number of Examples
D=1	Appendicitis (APP)	126
D=2	Diverticulitis (DIV)	28
D=3	Perforates Peptic Ulcer (PPU)	9
D=4	Non-Specific Abdominal Pain (NAP)	585
D=5	Cholecystitis (CHO)	53
D=6	Intestinal Obstruction (INO)	68
D=7	Pancreatitis (PAN)	11
D=8	Renal Colic (RCO)	60
D=9	Dyspepsia (DYS)	173

4 Methodology

Our goal is to find the most useful features' set for separating two kinds of diagnosis by using conformal predictors.

Firstly, we try to separate the two classes just by one feature. We take all examples from the dataset that belong to one of these two classes and do not contain missing values. Then we process it in on-line mode: prediction for a next example is based using all preceding ones as the training set, and then get the corresponding confidence of the single prediction. To assess performance we calculate average confidence of the examples, and Due to on-line processing, it is averaged over different sizes of the training set.

This was done for each feature, so the feature with the largest average confidence will be the first important feature of useful features' set.

Then we solve in the same way the question: what the second feature can be added to this one in order to maximize average confidence? After deciding this we will have list of two features, then we look for the third feature and so on.

On each step a feature is added to the list of ones being used for prediction, and average confidence grows until adding more features appears not to be useful anymore, the confidence does not change much. The speed of the method is depending on the size of examples and how many m suitable features we want from the full feature set.

5 Results and Discussion

We choose a typical result for illustration, and you can find more results at appendix. Table 1 shows the order we get when we separate APP (D1) from DYS (D9).

The order of features listed in tables corresponds to stage of including features into the selected set of features.

Table 1. Separate APP (D1)from DYS (D9)

Order	Value	Symptom	Average Confidence
1	4/2	Pain-site present: right lower quadrant	0.16
2	24/0 or 24/1	Abdominal scar: present or absent	0.32
3	5/3	Aggravating factors: food	0.44
4	6/5	Relieving factors: nil	0.48
5	14/0 or 14/1	Indigestion: history of indigestion or no history	0.64
6	27/0 or 27/1	Rebound: present or absent	0.73
7	9/0	Type of pain: steady	0.81
8	26/2	Site of tenderness: right lower quadrant	0.86
9	33/4	Rectal examination: normal	0.88
10	2/1	Age: 10-19	0.91

Figure 1 shows us the tendence of confidence average while separate APP(D1) from DYS(D9) when we extend features size till take every feature into account. As we can seen from figure, there are just 102 features, because when we calculate for useful features, some features show same values as equal useful features and we just choose one of them for further steps. The confidences in this picture grows fast at the beginning till meet the peak, 0.97503, and then keep steady for a while, near the end part, it get a slight fall to 0.958. This kind of fall also happens in other separation cases we mentioned above and may cause by the influence from features which are not relevant with this separation.

And, the programme finish its learning process when confidence reached 0.95 with the useful feature size is around 16 out of 135 features and then reach a plateau.

Figure 1 shows that the confidence level is greater than 0.95 when useful features'size is 20, so we will compare the results of using these 20 features with using whole features for separation D1 from D9 by on-line conformal prediction, and significance level ϵ is 0.05.

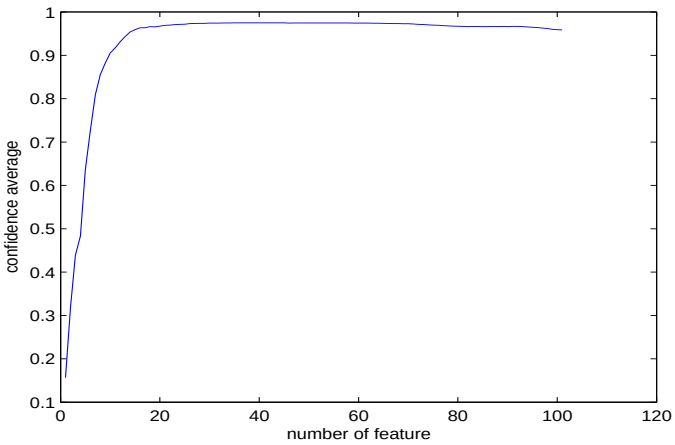

Fig. 1. Separate APP(D1) from DYS(D9)

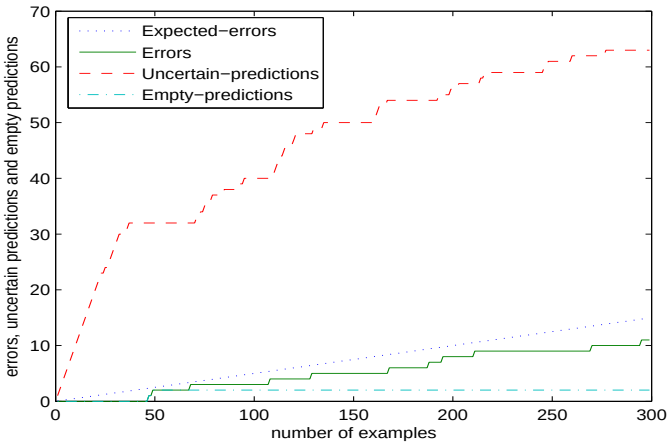

Fig. 2. Use the whole feature set to separate APP(D1) from DYS(D9)

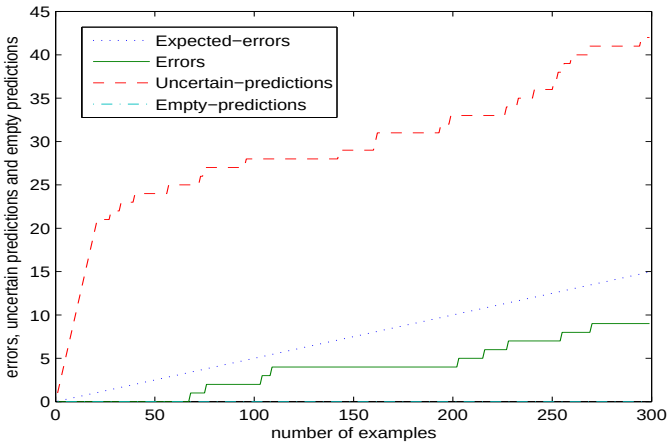

Fig. 3. Use 20 useful features to separate APP(D1) from DYS(D9)

Figure 2 presents the results of using all the features - we have 11 errors, 63 uncertain predictions and 2 empty predictions. Figure 3 shows the prediction results using only 20 selected features. It is clear that the number of prediction errors is reduced to 9 and we just get 42 uncertain predictions and 0 empty prediction. Thus, the selected 20 features give us better prediction results with less errors and more efficient size.

The following Table 2 shows us the comparison of results between full features and selected features in other binary classification subproblems by on-line conformal prediction. Compare with using all features, selected features could give us the same level of accuracy by small size. Because the predictions are under the guarantee, for a specific significant level, the best result is which has the most efficient prediction region. We will find how many features could provide us the best results which have the least uncertain predictions when significance level ϵ is 0.05.

Table 2. Results Comparison

	full features			selected features		the best results	
subproblem	size	accuracy	uncertain predictions	size	accuracy	size	uncertain predictions
D1–D9	135	0.96	63	17	0.96	30	31
D2–D9	135	0.98	76	22	0.98	26	30
D3–D9	135	0.97	41	20	0.97	21	26
D5–D9	135	0.96	173	17	0.96	40	92
D6–D9	135	0.97	101	14	0.97	33	40
D8–D9	135	0.96	69	18	0.96	30	44

6 Conclusion

As we can seen from above tables, at beginning, confidences are always low because few features are not enough for accurate predictions. As the number of features growing, the corresponding average confidence increases, and we could get desirable confidences by small number of features . One can then use a significance level in order to decide where to stop adding them. If a plateau is not reached, one can stop when confidence stops to grow, but this would probably mean that the underlying method of computing non-conformity measure was not very appropriate. So, conformal predictor could be an useful way of feature selection.

Acknowledgements. This work was supported in part by funding from BBSRC for ERASySBio+ Programme: Salmonella Host Interactions PRoject European Consortium (SHIPREC) grant; VLA of DEFRA grant on Development and Application of Machine Learning Algorithms for the Analysis of Complex Veterinary Data Sets; EU FP7 grant O-PTM-Biomarkers (2008–2011); and by grant PLHRO/0506/22 (Development of New Conformal Prediction Methods with Applications in Medical Diagnosis) from the Cyprus Research Promotion Foundation.

References

1. T. Bellotti, Z. Luo and A. Gammerman. Strangeness Minimisation Feature Selection with Confidence Machines, IDEAL 2006, Lecture Notes in Computer Science 4224, pp. 978-985, 2006.
2. I. Guyon and A. Elisseeff, *Journal of Machine Learning Research***3**, pp. 1157–1182, 2003.
3. J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.
4. F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65: pp 386–408, 1959.
5. V. Vovk, A. Gammerman and G. Shafer. *Algorithmic Learning in a Random World*, Springer, 2005.
6. Glenn Shafer, Vladimir Vovk. A Tutorial on Conformal Prediction, Journal of Machine Learning Research 9(2008) 371-421.
7. A. Gammerman and A.R. Thatcher. Bayesian Diagnostic Probabilisties without Assuming Independence of Symptoms. *Methods Inf Med.*. 30(1), pp. 15-22, 1991.
8. A. Gammerman and V. Vovk. Hedging Predictions in Machine Learning. The computer Journal, Vol. 50, No.2, pp. 151-163, 2007.
9. K. Proedrou, I. Nouretdinov, V. Vovk, and A. Gammerman. Transductive confidence machines for pattern recognition. In Proceedings of the 13th European Conference on Machine Learning (ECML'02), volume 2430 of LNCS, p.p. 381-390. Springer, 2002
10. H. Papadopoulos, A. Gammerman and V. Vovk. Reliable Diagnosis of Acute Abdominal Pain with Conformal Prediction. Engineering Intelligent Systems 17(2-3): 127-137. CRL Publishing. 2009.

Appendix

Table 3 is for separating of classes DIV(D2) and DYS(D9), Table 4 is for CHO(D5)and DYS(D9) and Table 5 is for PPU(D3) and NAP(D4).

Table 3. separate DIV (D2)from DYS (D9)

order	value	symptom	average confidence
1	4/12	Pain-site present: epigastric	0.27
2	2/2	Age: 20-29	0.43
3	2/3	Age: 30-39	0.54
4	20/0 or 20/1	Drugs: being taken or not being taken	0.54
5	26/13	Site of tenderness: none	0.67
6	16/1	bowel habit: constipated	0.72
7	10/0 or 10/1	Severity of pain: moderate or severe	0.76
8	6/0	relieving factors: lying still	0.82
9	21/0	mood: normal	0.83
10	22/1	color: pale	0.86
11	3/5	Pain-site onset: lower half	0.88
12	3/3	Pain-site onset: left lower quadrant	0.90
13	11/0 or 11/1	Nausea: nausea present or no nausea	0.91
14	8/0	Duration of pain: under 12 hours	0.92
15	14/0 or 14/1	Indigestion: history of indigestion or no history	0.93
16	3/4	Pain-site onset: upper half	0.932
17	8/1	Duration of pain: 12-24 hours	0.938
18	4/5	Pain-site present: lower half	0.941

Table 4. separate CHO (D5)from DYS (D9)

order	value	symptom	average confidence
1	22/3	Color: jaundiced	0.06
2	31/0 or 31/1	Murphy's test: positive or negative	0.10
3	24/0 or 24/1	Abdominal scar: present or absent	0.18
4	6/0	Relieving factors: lying still	0.25
5	18/0 or 18/1	Previous pain: similar pain before or no pain before	0.32
6	1/0 or 1/1	sex: male or female	0.45
7	4/12	Pain-site present: epigastric	0.59
8	5/5	Aggravating factors: nil	0.64
9	26/0	Site of tenderness: right upper quadrant	0.69
10	10/0 or 10/1	Severity of pain: moderate or severe	0.76

Table 5. Separate PPU(D3) from NAP(D4)

order	value	symptom	average confidence
1	2/1	Age: 10-19	0.25
2	10/0 or 10/1	Severity of pain: moderate or severe	0.54
3	26/2	Site of tenderness	0.67
4	2/2	Age: 20-29	0.76
5	26/13	Site of tenderness: none	0.83
6	18/0 or 18/1	Previous pain: similar pain before or no pain before	0.87
7	22/0	Color: normal	0.89
8	3/8	Pain-site onset: central	0.92
9	21/2	Mood: anxious	0.93
10	8/2	Duration of pain: 24-48 hours	0.94
11	2/5	Age: 50-59	0.945
12	4/2	Pain-site present: right half or left half	0.946