Feature Selection Using Non Linear Feature Relation Index

Namita Jain and C.A. Murthy

Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India
namita.saket@gmail.com,
murthy@isical.ac.in
http://www.isical.ac.in/~murthy/

Abstract. In this paper we propose a dependence measure for a pair of features. This measure aims at identifying redundant features where the relationship between the features is characterized by higher degree polynomials. An algorithm is also proposed to make effective use of this dependence measure for the feature selection. Neither the calculation of dependence measure, nor the algorithm need the class values of the observations. So they can be used for clustering as well as classification.

1 Introduction

The problem of reducing the number of features is also known as *dimensionality reduction*. Mathematically the problem of dimensionality reduction can be stated as follows. Given a p-dimensional random variable $x = (x_1, x_2, \cdots, x_p)$, find a lower dimensional representation of data $s = (s_1, \cdots, s_k)$ with $k \leq p$, that captures the content of the original feature set according to some criterion [1]. The two approaches used for dimensionality reduction are feature selection and feature extraction. Feature extraction aims at building a transformed feature set, such that the new feature set has some advantages over the original feature set. Principal Component Analysis is an example of a feature extraction technique. Feature selection aims at finding a representative feature set from the original set. Here the new representative set is the subset of original feature set. Irrelevant features are the ones which can be removed without making a negative impact on the learning task. Redundant features are features which become irrelevant due to presence of other features.

1.1 Feature Selection

Two important steps of feature selection are feature search and evaluating the feature subset according to some criterion [2][3][4].

Feature Search. Several search algorithms are available for feature selection in literature. None of them guarantees to provide the optimal feature subset for any choice of criterion function. Some of the search methods are branch

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 7–12, 2009.

and bound (performs poorly on non monotonic criterion functions), Sequential Forward search, Sequential Backward search, Sequential forward floating search, Sequential backward floating search [5]. Some algorithms use randomized search [1].

Evaluation criteria. Once a subset is generated, an evaluation criterion is used to assign a goodness level to a given subset of features. The criterion function can take different aspects into consideration. These include (i) Distance measure, (ii) Information measure, (iii) Dependence measure, (iv) Consistency measure [4].

Dependence Measure. Even if two features are individually relevant, the presence of one feature may make the other feature dispensable. Such features can be identified using a dependence measure [4]. Identifying such features is the main purpose of the criterion function defined in this paper.

Filter approach to feature selection. A feature selection method may follow either a filter approach or a wrapper approach[3]. The wrapper approach uses a learning algorithm that will ultimately be applied, as its guide, for selecting the features. The filter approach however uses the data alone to select the features. For an algorithm depending on filter approach, the evaluation of the feature set is not done in reference with any particular classifier. If the criterion function, used to determine the features to be added or removed to the subset, considers the effect of the change on entire subset rather then evaluating features separately, the whole set is evaluated at each step.

2 Proposed Feature Relation Index

In this section we propose a feature selection index which aims at identifying statistical dependence between features. A few existing dependence measures are also discussed here.

Correlation coefficient(| ρ |). A simple measure of relationship between two variables x and y is absolute value of correlation coefficient [6]. The correlation coefficient is given by given by $\rho(x,y) = \sum \frac{(x_i-\overline{x})(y_i-\overline{y})}{n\sqrt{var(x)var(y)}}$ where $\overline{x}$ and $\overline{y}$ are mean values of the two variables. The magnitude of $\rho(x,y)$ determines the strength of the relationship.| $\rho(x,y)$ | is invariant to translation and scaling. | $\rho(x,y)$ | is not rotation invariant [6].

Least Square Regression Error (e). Another measure which has been used as a measure of dependence between variables is Least Square error [6]. A straight line $y = a+bx$ is fit to a set of points $(x_i, y_i), i = 1, \cdots, n$ in the x, y plane, such that it minimizes the mean square error given by $e(x,y) = \frac{1}{n}\sum(e(x,y)_i)^2$, where $e(x,y)_i = y_i - a - bx_i$. The coefficients are given by $a = \overline{y} - b\overline{x}$ and $b = \frac{cov(x,y)}{var(x)}$. The mean square error is given by $e(x,y) = var(y)(1-\rho(x,y)^2)$. Mean square error is the residual variance unexplained by the linear model.

The value e denoting least square error (linear) will be 0 only if a perfect linear relationship exists between x and y. The measure is invariant to translation and sensitive to scaling. It is also non-symmetric. Having a non symmetric measure allows us to choose the variable which is more likely to have higher information content.

Maximal Information Compression Index (MICI)(λ_2)[7]. The linear dependence between two random variables x, y can be judged from smallest eigenvalue λ_2 of their covariance matrix $\sum$. The value of MICI is given by $2\lambda_2(x,y) = var(x) + var(y)\sqrt{(var(x)+var(y))^2 - 4var(x)var(y)(1-\rho(x,y)^2)}$ The measure λ_2 can be seen as the eigenvalue for the direction normal to the Principal Component Axis [17]. Therefore λ_2 indicates the minimum amount of error incurred on projecting the data on a single dimension.

The value of λ_2 will be 0 if and only if a perfect linear relationship exists between x and y. The measure is invariant to translation. It is sensitive to scaling. The measure is symmetric. The measure is invariant to rotation. This method will not be able to find non-linear realtionship between the variables.

2.1 Non-linear Feature Relation Index

Here, we try to approximate one variable using a polynomial of another variable. This is done in such a way, that error function is minimized. We try to predict the value of $\hat{y}$ as a polynomial function of x

$$\hat{y} = \beta_0 + \beta_1 x + \cdots + \beta_k x^k$$

The error function which is minimized is given by

$$e_{yx} = \sum (\hat{y} - y)^2$$

Thus, the value of e_{yx} is minimized to find the coefficients of the polynomial used to model the relationship between the variables. we obtain (k+1) equations for (k+1) unknowns.

For first degree polynomial this error is same as the least square error. Here the use of higher degree polynomilas allow us to discover relationships described by polynomial curves. Just as the error in that case was proportional to variance the error here is proportional to higher degree moments of the variable y.

The value of proposed measure e_{yx} will be 0 only if y can be represented perfectly as a polynomial function of x. The measure is invariant to translation. It is sensitive to scaling. The error function obtained is proportional to higher degree central moments of the variable y. Since, central moment is a good indicator of information content, this allows us to choose a variable with more information content. The measure is non symmetric. The measure is not invariant to rotation. The measure is able to find non-linear relationship between the variables.

2.2 Selecting a Value for k

In the given method we try to get a better estimation of y as a higher degree polynomial of x. This will lead to a smaller error value. It may be noted that

increasing the power of the polynomial will not adversely effect the performance of the algorithm even if y can be represented as a lower degree polynomial of x. Let us consider the case where y can be represented perfectly as a polynomial (of dgree k) of x, leading to zero error. Let the representation be $y = \beta_{k,0} + \beta_{k,1}x + \cdots + \beta_{k,k}x^k$. Now we try to represent y as a k+1 degree polynomial of x as $y = \beta_{k+1,0} + \beta_{k+1,1}x + \cdots + \beta_{k+1,k+1}x^{k+1}$. We find that the value of the coefficients we obtain are given by

$$\beta_{k+1,i} = \beta_{k,i} for 1 <= i <= k, \beta_{k+1,k+1} = 0$$

3 Proposed Feature Selection Algorithm

First step of this feature selection method is to calculate the value of Non linear feature relation index e_m for each pair of features. The objective is to iteratively select features for which the elimination of dependent variables will lead to minimum loss. We select a parameter r which indicates the maximum number of features which can be removed at each step. Select number of features to be removed in the next step as $n = min(r, c)$ where c represents the number of features left to be considered. For each feature the n pairs which have minimum value of e_m are selected and summation of e_m is calculated over these pairs. The feature for which the calculated sum is minimum is selected and the corresponding dependent features are eliminated at each step. These steps are repeated till all the features are included either in set of selected features or in the set of rejected features.

3.1 Algorithm

Let F be the original set of features, S be the set of selected features and E be the set of eliminated features. Choose a value k giving the degree of expectation function to be used to Regression. Choose r giving the maximum number of features to be eliminated at each step. Let ϵ be the maximum error caused by eliminating a feature.

1. For each feature $i \in F$
   (a) For each feature $j \in F$ $M(i, j) \leftarrow e_m(i, j)$
   (b) $\epsilon = \infty$
   (c) $r = min(r, card(F - (S \cup E)))$
   (d) Find the set of features $D_i = [j_1, \cdots, j_r]$ such that

$$\forall(a, b)\{(a \in D_i, b \in F - D_i) \Rightarrow M(i, a) \leq M(i, b)\}$$

   (e) if $M(i, r) > \epsilon$ then
      i. $r = r - 1$
      ii. goto 1.(c)
   (f) $S(i) = \sum M(i, j_k)$,where $k = 1, \cdots, r$

2. Find the feature i such that

$$\forall(i,j)\{i \in F-(S \cup E) \wedge j \in F-(S \cup E \cup \{i\}) \Rightarrow S(i) < S(j)\}$$

3. if $card(S) = 0$ then $\epsilon = M(i,r)$, where i is the feature found in step 2.
4. $S = S \cup \{i\}$, where i is the feature found in step 2.
5. $E = E \cup D_i$, where D_i is the set found in step 1.(d) corresponding to the feature i selected in step 3.
6. For each $d \in D_i$, where D_i is same as in step 4 $M(d,t) = \infty$ and $M(t,d) = \infty$ for all $t \in F$
7. If $card(F-(S \cup E)) > 0$ goto step 1.(c)

4 Results

The non Linear Feature relation index has been used by the algorithm given in previous section. We have tested this algorithm on several real life datasets [8] such as Iris, Cover Type (10 numeric attributes are used.), Ionosphere, Waveform, Spambase, Isolet, and Multiple features. A polynomial of degree 3 has been used for curve fitting while finding the non Linear Feature relation index.

As per three dimensional framework for classification given by Huan Liu and Lei Yu, the proposed method falls in the category of dependence measure, using a sequential search strategy, useful for clustering task[3] along with MICI. The performance of the proposed algorithm has been compared to the results given by MICI [7][9]. The proposed method gives better performance for all the datasets except Isolet and Spambase. It can be seen that the proposed method results in a slightly increased error rate for Spambase and Isolet datasets. In Isolet dataset the data is given in a normalized form. In Spambase most of the features are scaled to be represented as percentage. Scaling of data might be a reason for reduced performance of the proposed algorithm in the two cases as the Non Linear Feature Relation Index is sensitive to scaling. The difference in performance has been found to be statistically significant in all the datasets using the Welch test for Behrens-Fisher problem.

Table 1. Classification performance of proposed feature selction algorithm as compared to MICI, where D denotes the number of features in original set and d_{MICI} and d denote the number of features in the reduced set obtained by MICI and the proposed method

Dataset	D	d_{MICI}	MICI	d	Proposed
Iris	4	2	97.33	2	97.47
Cover Type	10	5	63.55	5	66.56
Ionosphere	32	16	65.92	11	80.57
Waveform	40	20	63.01	20	87.94
Spambase	57	29	88.19	29	87.47
Isolet	610	310	95.01	307	94.65
Mfeat	649	325	78.34	324	94.11

5 Conclusion

In this paper we proposed a dependence measure based on polynomial regression between two features. Also, an algorithm is proposed which effectively uses this measure to identify the redundant features from a given set of features. Since the calculation of dependence measure and the algorithm are independent of the class value of the observations, the proposed method is suitable for clustering tasks along with classification. The algorithm is used to obtain reduced feature sets for real life datasets. The reduced feature sets are then used for classification task. A good classification performance here indicates that redundant features have been discarded from the original feature sets.

References

1. Liu, H., Motoda, H.: Computational Methods of Feature Selection. Chapman Hall/CRC
2. Dash, M., Liu, H.: Feature Selection for Clustering. In: Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications, April 18-20, 2000, pp. 110–121 (2000)
3. Liu, H., Yu, L.: Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Transactions on Knowledge and Data Engineering 17(4), 491–502 (2005)
4. Dash, M., Liu, H.: Feature Selection for Classification. Intelligent Data Analysis 1(3), 131–156 (1997)
5. Devijver, P.A., Kittler, J.: Pattern Recognition: A Statistical Approach. Prentice Hall, Englewood Cliffs (1982)
6. Hoel, P.G., Port, S.C., Stone, C.J.: Introduction to Statistical Theory. Houghton Mifflin, New York (1971)
7. Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised Feature Selection Using Feature Similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(3), 301–312 (2002)
8. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository, `http://www.ics.uci.edu/~mlearn/MLRepository.html`
9. Pal, S.K., Mitra, P.: Patter Recognition Algorithms for Data Mining. ChapMan and Hall/CRC