Multiple-Side Multiple-Learner for Incomplete Data Classification

Yuan-ting Yan, Yan-Ping Zhang(✉), and Xiu-Quan Du

Key Laboratory of Intelligent Computing and Signal Processing of Ministry
of Education, School of Computer Science and Technology,
Anhui University, Hefei 230601, Anhui Province, China
365975632@qq.com, zhangyp2@gmail.com

Abstract. Selective classifier can improve classification accuracy and algorithm efficiency by removing the irrelevant attributes of data. However, most of them deal with complete data. Actual datasets are often incomplete due to various reasons. Incomplete dataset also have some irrelevant attributes which have a negative effect on the algorithm performance. By analyzing main classification methods of incomplete data, this paper proposes a Multiple-side Multiple-learner algorithm for incomplete data (MSML). MSML first obtains a feature subset of the original incomplete dataset based on the chi-square statistic. And then, according to the missing attribute values of the selected feature subset, MSML obtains a group of data subsets. Each data subset was used to train a sub classifier based on bagging algorithm. Finally, the results of different sub classifiers were combined by weighted majority voting. Experimental results on UCI incomplete datasets show that MSML can effectively reduce the number of attributes, and thus improve the algorithm execution efficiency. At the same time, it can improve the classification accuracy and algorithm stability too.

Keywords: Incomplete data · Multiple-side · Feature subset · Multiple-learner

1 Introduction

When solving a problem, human usually ignore the irrelevant details and focus on the major part of the problem, in this way, they can simplify the problem solving. For example, feature selection [1], attribute reduction [2] in knowledge mining, etc. In addition, analyzing problem in several different aspects and then combing their results is another common solution of human problem solving. There are many related researches, such as subspace [3], multiple view learning [4], and so on.

These two ways of problem solving have been widely used in classification problem [5,6]. First of all, ignore irrelevant information can improve the algorithm execution efficiency. Studies have shown that irrelevant attributes have a negative effect on classification accuracy. Secondly, classifying from several

© Springer International Publishing Switzerland 2015

Y. Yao et al. (Eds.): RSFDGrC 2015, LNAI 9437, pp. 322–331, 2015.
DOI: 10.1007/978-3-319-25783-9_29

different views and then combine their results is another effective method to improve classification accuracy. However, most of the researches are deal with complete data. At the same time, in many practical applications, missing values are often inevitable due to various reasons. Such as equipment errors, data loss, manual data input, etc. So, classification on incomplete data is very necessary.

To avoid the negative impact of irrelevant attributes on the classification performance, we propose a multiple-side multiple-learner algorithm (MSML) for incomplete data. MSML first uses chi-square statistic evaluation algorithm to delete some unimportant attributes, and then constructs a group of classifiers according to the missing feature values in the selected feature subset. Finally, the results of different classifiers are combined by weighted majority voting.

The rest of the paper is organized as follows. The research on incomplete data classification is briefly reviewed in Sect. 2. In Sect. 3, we introduce the MSML algorithm. Section 4 gives the numerical experiments on 8 real incomplete datasets form UCI Machine Learning Repository. Section 5 concludes the paper.

2 Related Work

Some scholars have studied the classification on incomplete data. There are two simple methods to deal with incomplete datasets. One way is simply ignore the samples with missing values. However, this may cause loss of potential profitable information, leading to an insufficient amount of samples for investigation [7]. Imputation method is another common solution to replace missing values with a particular value of the individual variables. Both methods are known to incur potential estimation bias [8,9]. One kind of methods can avoid the estimation bias is to use the EM algorithm [10], gradient descent [11], Gibbs sampling [12] or Logistic regression algorithm [13]. But this kind of methods relies on the assumption that data are missing at random and there is no technique to verify this assumption. Meanwhile, this kind of methods will suffer a dramatic decrease in accuracy when this assumption is violated.

To avoid the missing at random assumption, Ramoni and Sebastiani proposed a Robust Bayes Classifier (RBC) [14] that needs no assumption about data missing mechanism. However, similar to Naive Bayes Classifier, RBC also makes the assumption that attributes are independent for each class. Krause et. al [15] introduced an ensemble method to deal with incomplete data, sub classifiers were trained on random feature subsets. The method also assumed that the value of any feature is independent of all others. Chen et.al [16] put forward a noninvasive neural network ensemble (NNNE) method without any assumptions about the data distribution. This method generates a community of base classifiers trained only with known values. But it did not take into account the differences of attribute importance degree. To overcome the limitation, a multi-granulation ensemble method (MGNNE) was proposed [17]. Information entropy was applied to measure attribute importance degree. However, the performance of MGNNE relies on the proportion of samples whit no missing values. Moreover, all the above three methods did not consider the negative effect of irrelevant attributes on classification performance.

Considering the characteristic of incomplete dataset and the negative effect of irrelevant attributes on classification performance. We propose a new algorithm called multiple-side multiple-learner classification algorithm (MSML) to deal with incomplete data.

3 MSML

3.1 Chi-Square Statistic Feature Evaluation Algorithm

We apply chi-square statistic to calculate the importance degree between each attributes and class variable respectively. A feature subset is selected by removing the attributes with cumulative probability distribution (*cdf*) values smaller than threshold α. We first give the method to construct the contingency table of an attribute variable with respect to the class variable.

Given an incomplete dataset D, suppose A is an attribute of D with m values, d is the class variable with l values. Note, we use '?' to denote the missing (unknown) value. The process of constructing contingency table M_A of attribute A with respect to d can be described as follows:

(1) Count the following frequencies:
$f_{ij} = f(A = a_i, d = d_j), f_{(m+1)j} = f(A =?, d = d_j)$,
$f_{i(l+1)} = f(A = a_i, d =?)$ and $f_{(m+1)(l+1)} = f(A =?, d =?)$.

(2) Allocate $f_{(m+1)j}$, $f_{i(l+1)}$ and $f_{(m+1)(l+1)}$ to f_{ij}.
To updatef_{ij}:

(2.1) Compute the following summation:
$$row_i = \sum_{j=1}^{l} f_{ij}, col_j = \sum_{i=1}^{m} f_{ij}, N = \sum_{i=1}^{m}\sum_{j=1}^{l} f_{ij}$$
(2.2) Update f_{ij}:
$$f_{ij}^{'} \leftarrow f_{ij} + f_{i(l+1)} \times \frac{col_j}{N} + f_{(m+1)j} \times \frac{row_i}{N} + f_{(m+1)(l+1)} \times \frac{f_{ij}}{N}$$
(3) Obtain the contingency table $M(M_{ij} = f_{ij}^{'})$

According to the above steps, we can get all the contingency tables between each attribute and class variable, respectively. Given a contingency table M of attribute A with respect to d, we use the chi-square statistics to measure the importance degree of attribute A. The chi-square attribute evaluation algorithm of incomplete dataset is as follows:

(1) Construct the contingency table M $(m * n)$ of each attribute with respect to class variable, m and n are the number of distinct values (except '?') of attribute A and class variable d, respectively;
(2) For a contingency table M of attribute A with respect to d, compute the chi-square statistic $Chi(A, d)$;

(2.1) Compute the summation of each row of M_A denoted by r_i and each column of M_A denoted by c_j, respectively:
$r_i = \sum\limits_{j=1}^{n} v_{ij}, (i = 1, 2, ..., m), c_j = \sum\limits_{i=1}^{m} v_{ij}, (j = 1, 2, ..., n);$

(2.2) For each pair of (i, j), calculate the expected frequency E_{ij}:
$E_{ij} = \frac{r_i.c_j}{N}(i = 1, 2, ..., m; j = 1, 2, ...n; N = \sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n} f_{ij});$

(2.3) Compute the chi-square statistic value $Chi(A, d) = \sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n} \frac{(E_{ij}-v_{ij})^2}{E_{ij}};$

(2.4) Compute the *cdf* value P_A corresponding to $Chi(A, d)$.

(3) Select the feature subset S_1 consist of attributes with *cdf* value bigger than threshold $\alpha(0 < \alpha < 1)$.

According to the above method, we can get a feature subset S_1 of the incomplete dataset. S_1 consists of attributes with *cdf* value bigger than a given threshold α. In general, S_1 still have missing values. We will construct a group of classifiers on S_1.

3.2 Multiple-Side Multiple-Learner for Incomplete Data

Let $D = \{(x_i, y_i)\,|i = 1, 2, ..., n\}$ be the incomplete dataset. Where n denote the size of the dataset. Suppose there are d features of the input space $X = (X^{(1)}, X^{(2)}, ..., X^{(d)})$. If a value $x_i^{(j)}$ of sample x_i is unknown, it is denoted as $x_i^{(j)} = null$. For convenience, we first give some definitions as follows:

Definition 1: For a sample x_i of D, the missing value set of sample x_i is defined as a feature subset $mset\{i\}$ that x_i is missing for all features in $mset\{i\}$ and is complete for all features in X but not in $mset\{i\}$.

$mset\{i\} = \{X^{(j)}\,|(\forall X^{(j)} \in mset\{i\} \wedge \mathrm{x}_i^{(j)} = null) \wedge (\forall X^{(j)} \notin mset\{i\} \wedge x_i^{(j)} \neq null)\}$.

Definition 2: The missing attribute set (MS) of D is defined as a set of missing value sets, $MS = \{MS_1, ..., MS_k\}$, in which each missing value set is unique.

Definition 3: A complete data subsets X_{mset_R} is defined as a data subset corresponding to a missing attribute set $mset_R$.

$X_{mset_R} = \{x_i^{(j)}\,|x_i \in D \wedge \forall j \notin mset_R(x_i^{(j)} \neq null)\}$

Note, each complete data subset corresponding to a unique feature subset (or missing attribute set) of the incomplete dataset.

To improve the algorithm performance, each complete data subset is used to train a classifier based on bagging algorithm. For a test sample, the algorithm chooses the classifiers that did not require the missing value of the test sample to predict it. And then weighted majority voting is applied to combine the prediction results of the test sample.

Algorithm 1. Multiple-side Multiple-learner Classification.

```
Input: Training dataset $D_{train} = \{(x_i, y_i)|i = 1, ..., n\}$.
  Testing dataset $D_{test} = \{(x_i, y_i)|i = 1, ..., m\}$.
Output: Prediction results $Y = \{Y_1, ..., Y_m\}$.
  Initialize $Y \leftarrow \emptyset$, $temp \leftarrow \emptyset$
  Training
  Obtain the missing attribute set $MS = \{MS_1, ..., MS_k\}$ and the complete data
  subsets $X = \{X_1, ..., X_k\}$ of $D_{train}$.
  Calculate the mutual information $MI = \{MI_1, ..., MI_k\}$.
  for $i \leftarrow 1$ to $length(X)$ do
      Generate a classifier $\mathbf{h_i}$ on $X_i$ by using bagging algorithm and bp network.
  end
  Testing
  for $j \leftarrow 1$ to $m$ do
      Obtain the missing value set $mset\{j\}$ of sample $j$ and set $temp \leftarrow \emptyset$;
      for $i \leftarrow 1$ to $length(MS)$ do
          if $mset\{j\} \subseteq MS_i$ then
              $temp = [temp, \mathbf{h_i}(x_j)]$;
          end
      end
  Obtain final result $Y_j$ of sample $j$ by using weighted majority voting.
  $Y \leftarrow [Y, Y_j]$.
  end
  return $Y$
```

In this paper,to determine the final prediction of test sample, some factors are concerned to realize the weighted majority voting. First, each complete data subset has a unique feature subset with an relevance degree for prediction the class label. Moreover, the sub classifiers trained on complete data subsets have different prediction accuracies, as is commonly agreed that higher testing accuracy tends to have greater prediction accuracy. Besides, the size of complete data subsets are different, it is also a factor need to be considered. Combining these three factors, each available sub classifier is assigned a weight by the following method.

$$w_i = \frac{MI_i \left| X_{mset_i} \right| ACC_i}{\sum MI_i \left| X_{mset_i} \right| ACC_i} \tag{1}$$

Here $|X_{mset_i}|$ denote the size of complete data subset X_{mset_i}, MI_j denote the relevance degree (measured by mutual information) between attributes set and class variable on data subset X_{mset_i}, ACC_i denote the testing accuracy of the i_{th} sub classifier. Algorithm 1 gives the MSML algorithm.

4 Experiments

4.1 Experimental Description

To testify the validity of MSML, we carried out experiments on 8 benchmark datasets with missing data from UCI machine learning repository [18]. All our experiments were programming by MatlabR2001a. The implementation was performed on an Intel Core i5 CPU running at 3.2GHz (4CPUs) and 4GB RAM. Table 1 gives the detail information about the datasets used for experiments.

For MSML, MGNNE and NNNE, a faster BP algorithm called Levenberg-Marquardt algorithm which has an efficient implementation provided by Matlab is used in our experiments. The number of input nodes (*id*) is determined by the number of available attributes on each data subsets, and the number of output nodes (*od*) is determined by the number of classes. According to the geometric pyramid rule, the number of hidden nodes is $\sqrt{id * od}$. We evaluate the accuracy using ten-folds cross validation approach where a given dataset is randomly partitioned into ten folds of equal size. For each complete data subset, we apply the bagging algorithm to improve the algorithm performance, and set 10 as the number of replicates [19].

Table 1. Summarization of datasets characteristics

Dataset name	Instance	Attributes	Classes
Automobile	205	26	6
Bands	540	39	2
B.cancer	699	10	2
Credit	690	15	2
Heart-h	294	13	2
Vote	435	16	2
L.cancer	32	56	2
Mushroom	8124	22	2

4.2 Experimental Results and Analysis

In our algorithm, the attributes with *cdf* values smaller than threshold α will be deleted to avoid the adverse effect of irrelevant attributes on algorithm performance. We choose two datasets Bands and L.cancer to study the relationship between algorithm performance and the threshold. We set the threshold to vary from 0.50 to 0.95 with the interval 0.05. Table 2 and Table 3 report the results.

One can see that, with the increase in the number of α, both the number of selected attributes and the algorithm runtime decreased gradually. During the process of α increased from 0.5 to 0.9, algorithm accuracy is basically unchanged. When $\alpha = 0.95$, the algorithm performance on both datasets has an obvious decline (Bands: about 2 % decline, L.cancer: about 46 % decline). From the experimental results, in this paper, we choose $\alpha = 0.9$ as the threshold to delete unimportant attributes, and thus to improve algorithm efficiency.

Table 4 gives the accuracy comparison of our algorithm to MGNNE, NNNE and RBC on 8 datasets. We can see that, overall speaking, NNNE has relatively poor performance. RBC has best accuracy on two datasets B.cancer and

Table 2. Performance of MSML on Bands with the change of α

α	#.Attributes	Accuracy	Runtime(s)
0.50	36	0.769	976
0.55	34	0.775	971
0.60	33	0.776	968
0.65	33	0.778	968
0.70	32	0.778	965
0.75	32	0.776	965
0.80	30	0.776	763
0.85	29	0.779	752
0.90	26	0.779	732
0.95	24	0.758	717

Table 3. Performance of MSML on L.cancer with the change of α

α	#.Attributes	Accuracy	Runtime(s)
0.50	44	0.577	11.2
0.55	41	0.574	10.5
0.60	38	0.579	10.4
0.65	34	0.575	5.1
0.70	32	0.574	5.1
0.75	28	0.573	4.9
0.80	25	0.579	4.9
0.85	23	0.573	4.9
0.90	17	0.576	4.9
0.95	11	0.300	0.5

Heart-h. MSML has best performance on 5 datasets, and MGNNE has a slightly better accuracy than MSML on dataset Vote. It indicates that there are a small amount of relevant attributes been removed from dataset Vote when we set $\alpha = 0.9$. One effective solution is to increase the number of selected attributes by setting a smaller threshold. On four datasets Automobile, Bands, Credit and L.cancer, compared with MGNNE, MSML has a certain improvement on accuracy ($1\,\% \sim 2\,\%$). It suggests that the irrelevant attributes has an adverse impact on algorithm accuracy.

By deleting irrelevant attributes, compared with NNE-based algorithms, the execution efficiency of MSML is greatly improved. Table 5 shows the details of three algorithms MSML, MGNNE and NNNE on 8 datasets. Note that the difference between MGNNE and NNNE is that MGNNE modified the weighted majority voting method of NNNE by applying information entropy to measure

Table 4. The average accuracy of the four classifiers

Datasets	MSML	MGNNE	NNE	RBC
Automobile	**70.05 ± 0.11**	68.18 ± 0.08	66.31 ± 0.10	68.49 ± 3.84
Bands	**77.85 ± 0.06**	75.62 ± 0.05	74.63 ± 0.06	71.36 ± 0.48
B.cancer	94.99 ± 0.02	93.99 ± 0.02	93.85 ± 0.02	**97.11 ± 0.16**
Credit	**86.45 ± 0.04**	85.50 ± 0.04	84.81 ± 0.06	86.18 ± 0.40
Heart-h	80.91 ± 0.07	80.59 ± 0.06	81.69 ± 0.06	**85.88 ± 2.11**
Vote	94.47 ± 0.02	**94.71 ± 0.03**	0.942 ± 0.03	90.25 ± 0.19
L.cancer	**57.83 ± 0.28**	52.17 ± 0.27	49.75 ± 0.28	56.13 ± 1.67
Mushroom	**99.96 ± 0.01**	99.96 ± 0.01	99.86 ± 0.01	95.96 ± 0.02

the importance degree of each sub classifiers. So the number of attributes and the number of data subsets of both methods are equal. Thus, we just list the runtime of MGNNE.

We can see that quite a few irrelevant attributes was deleted on three datasets Bands, Credit and Heart-h, so the number of complete data subsets decreased a lot, thus the algorithm computational time is greatly reduced. At the same time, the runtime of MSML is higher than MGNNE and NNNE on the two datasets Automobile and Mushroom. That is because both datasets has only one attribute was removed, and the number of data subsets is unchanged. However, the chi-square statistic attribute evaluation algorithm is introduced, which increases a certain algorithm execution time. For dataset L.cancer, the algorithm runtime has an apparent decline because its attributes number reduced from 56 to 17. Meanwhile, one can see that there is only one data subsets of MSML, which means that all the attributes with missing values are deleted. Overall, by removing irrelevant attributes, MSML can effectively enhance execution efficiency on the basis of guarantee algorithm accuracy.

Table 5. Runtime, number of selected attributes and number of data subsets

Dataset	Runtime			#.Subsets		#.Attributes	
	MSML	MGNNE	NNNE	MSML	NNNE	MSML	NNNE
Automobile	89.5	80.8	77.1	6	6	24	25
Bands	731.6	989	989	40	66	26	39
B.cancer	50.6	111	110.9	2	2	9	10
Credit	77.8	183	183	4	8	11	15
Heart-h	46	123.9	122.5	4.9	12	7	13
Vote	899	961	961	64.7	73	15	16
L.cancer	4.9	19.3	18.9	1	3	17	56
Mushroom	718.6	640.6	636	2	2	21	22

5 Conclusion and Discussion

By removing the irrelevant attributes of dataset, and then building ensemble classifier on the selected attributes set is an effective way to improve algorithm accuracy and execution efficiency. Most current studies require complete data. However, actual datasets are mostly incomplete due to various reasons, thus build classifier can deal with incomplete data is meaningful.

This paper puts forward a multiple-side multiple-learner classification algorithm to deal with incomplete data based on the characteristics of incomplete dataset. MSML first construct the contingency table of all attributes with respect to class variable, and then MSML introduces chi-square statistic evaluation algorithm to select a feature subset by removing the irrelevant attributes. Experiments show that MSML is an effective classification method to deal with incomplete dataset.

Acknowledgments. This work was supported by National Natural Science Foundation of China (Nos.61175046 and 61203290).

References

1. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
2. Qian, Y., Liang, J., Pedrycz, W., Dang, C.: Positive approximation: an accelerator for attribute reduction in rough set theory. Artif. Intell. **174**(9), 597–618 (2010)
3. Kuncheva, L.I., Rodrguez, J.J., Plumpton, C.O., et al.: Random subspace ensembles for fMRI classification. IEEE Trans. Med. Imaging **29**(2), 531–542 (2010)
4. Zhang, J., Zhang, D.: A novel ensemble construction method for multi-view data using random cross-view correlation between within-class examples. Pattern Recogn. **44**(6), 1162–1171 (2011)
5. Sun, S., Zhang, C.: Subspace ensembles for classification. Phys. A Stat. Mech. Appl. **385**(1), 199–207 (2007)
6. Bryll, R., Gutierrez-Osuna, R., Quek, F.: Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. Pattern Recogn. **36**(6), 1291–1302 (2003)
7. Allison, P.D.: Missing Data. Sage Publications, Thousand Oaks (2001)
8. Roderick L., J A, Rubin, D.B.: Statistical Analysis with Missing Data, vol. 43, no. 4, pp. 364–365. Wiley, New York (2002)
9. Gheyas, I.A., Smith, L.S.: A neural network-based framework for the reconstruction of incomplete data sets. Neurocomputing **73**(16), 3039–3065 (2010)
10. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. Ser. B (Methodol.) **39**, 1–38 (1977)
11. Russell, S., Binder, J., Koller, D., Kanazawa, K.: Local learning in probabilistic networks with hidden variables. In: Proceedings of IJCAI 1995, pp. 1146–1152 (1995)
12. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721–741 (1984)

13. Williams, D., Liao, X., Xue, Y., Carin, L., Krishnapuram, B.: On classification with incomplete data. IEEE Trans. Pattern Anal. Mach. Intell. **29**(3), 427–436 (2007)
14. Ramoni, M., Sebastiani, P.: Robust Bayes classifiers. Artif. Intell. **125**(1), 209–226 (2001)
15. Krause, S., Polikar, R.: An ensemble of classifiers approach for the missing feature problem. In: IEEE Proceedings of the International Joint Conference on Neural Networks, vol. 1, pp. 553–558 (2003)
16. Chen, H., Du, Y., Jiang, K.: Classification of incomplete data using classifier ensembles. In: IEEE International Conference on Systems and Informatics. pp. 2229–2232 (2012)
17. Yan, Y.-T., Zhang, Y.-P., Zhang, Y.-W.: Multi-granulation ensemble classification for incomplete data. In: Miao, D., Pedrycz, W., Slezak, D., Peters, G., Hu, Q., Wang, R. (eds.) RSKT 2014. LNCS, vol. 8818, pp. 343–351. Springer, Heidelberg (2014)
18. UCI Repository of machine learning databases for classification. http://archive.ics.uci.edu/ml/datasets.html
19. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)