SCIENCE CHINA
Information Sciences

• **RESEARCH PAPERS** •

January 2011 Vol. 54 No. 1: 1–13
doi: 10.1007/s11432-011-4406-5

Symptom selection for multi-label data of inquiry diagnosis in traditional Chinese medicine

SHAO Huan[1], LI GuoZheng[2]*, LIU GuoPing[3] & WANG YiQin[3]

[1] *School of Computer Engineering and Science, Shanghai University, Shanghai* ~~20072~~*, China;*
[2] *Department of Control Science and Engineering, Key Laboratory of Ministry of Education for Service Computing and Embedded Systems, Tongji University, Shanghai* 201804*, China;*
[3] *Laboratory of Information Access and Synthesis of TCM Four Diagnosis, Shanghai University of Traditional Chinese Medicine, Shanghai* 201203*, China*

Received August 18, 2010; accepted December 22, 2010

Abstract In traditional Chinese medicine (TCM) diagnosis, a patient may be associated with more than one syndrome tags, and its computer-aided diagnosis is a typical application in the domain of multi-label learning of high-dimensional data. It is common that a great deal of symptoms can occur in Traditional Chinese Medical diagnosis, which affects the modeling of diagnostic algorithm. Feature selection entails choosing the smallest feature subset of relevant symptoms, and maximizing the generalization performance of the model. At present there are rare researches on feature selection on multi-label data. A hybrid optimization technique is introduced to symptom selection for multi-label data in TCM diagnosis in this paper, and modeling is made by means of four multi-label learning algorithms like k nearest neighbors, etc. We compare the performance of the algorithm with the current popular dimension reduction algorithms like MEFS (Embedded Feature Selection for Multi-Label Learning), MDDM (Multi-label Dimensionality reduction via Dependence Maximization) on the UCI Yeast gene functional data set and an inquiry diagnosis dataset of coronary heart disease (CHD). Experimental results show that the algorithm we present has significantly improved the performance. In particular, the improvement on the average_precision for the classifier is up to 10.62% and 14.54%. Syndrome inquiry modeling of CHD in TCM is realized in this paper, providing effective reference for the diagnosis of CHD and analysis of other multi-label data.

Keywords multi-label learning, feature selection, high-dimensionality, inquiry of Traditional Chinese Medicine, coronary heart disease

Citation Shao H, Li G Z, Liu G P, et al. Symptom selection for multi-label data of inquiry diagnosis in traditional Chinese medicine. Sci China Inf Sci, 2011, 54: 1–13, doi: 10.1007/s11432-011-4406-5

1 Introduction

It is an effective way to promote TCM to information by applying the machine learning techniques to extract information in clinical experience to achieve the summarization and heritage of famous doctor [1]. In TCM clinical data, a case has many symptoms and may be associated with more than one syndrome. This is a typical analysis problem of high-dimensional multi-label data.

If a sample is associated with more than one class of labels, we call data like this multi-label data. Multi-label learning tasks are omnipresent in real-world problems. For instance, in text categorization, each

*Corresponding author (email: gzli@tongji.edu.cn) document may belong to several predefined topics, such as government and health; in scene classification each image may belong to several semantic classes, such as beach and urban. In all these cases, each instance is associated with a set of labels, and the task is to output a label set whose size is unknown a priori for each unseen instance. The existing techniques can be divided into two categorizations [2]: problem transformation methods and algorithm adaption methods. Problem transformation methods transform the learning task into one or more single-label classification tasks and they are algorithm-dependent. Some could be used for feature selection as well. Algorithm adaption methods extend specific learning algorithms (like SVM, decision tree, and neural network) to handle multi-label data directly.

Multi-label learning is usually referred to high-dimensional data, but there are very few dimension reduction methods and feature selection methods available for multi-label data due to the complexity of multi-label learning. As to feature dimensionality reduction, the recently published MDDM (multi-label dimensionality reduction via dependence maximization) [3] is a feature extraction method which uses the HSIC as the performance criteria and attempts to project the original data into a lower-dimensional feature space to maximize the dependence of the original feature description on the associated class labels. Experiments show that MDDM is slightly superior to principal component analysis PCA and nonlinear dimensionality reduction method LPP, and is significantly superior to the multi-label dimensionality reduction method MLSI [4]. Linear dimensionality reduction [5] shows improved performance when the least squares and other loss functions, including the hinge loss and the squared hinge loss, are used in multi-label classification. One problem of MDDM and linear dimensionality reduction is that the original low-dimensional features cannot be obtained, which poses an obstacle to scientific understanding of scientific problems.

Feature selection attempts to remove irrelevant and redundant features and entails choosing the smallest number of features to adequately represent the data and maximizing the prediction or classification accuracy. Feature selection distinctly improves the comprehensibility of the classification model and builds a model which can better predict the unknown samples. It has practical significance. For example, extensive experiences are needed to grasp the main symptom in TCM differential treatment. The current feature selection methods are divided into three broad categories: wrappers, filters, and embedded methods [6]. Wrappers depend on the learning machine and utilize the learning machine of interest as a black box to score feature subsets according to their predictive power. Although the wrapper methods are comparatively time-consuming, they are widely used in scientific data analysis because the selected feature subset is optimal to the specific learning machine due to its mechanism that the selection result is based on the learning algorithms. In multi-label feature selection, MEFS (embedded feature selection for multi-Label learning) [7] was proposed last year, in which sequential backward search algorithm is adopted to search the feature subset, and the prediction risk criterion [8] is used to evaluate the performance of the feature subset. In wrappers, a comparatively good result was achieved when the genetic algorithm was introduced [9]. A hybrid optimization technique which combines several optimization techniques to improve performance of multi-label learning is proposed in this paper, and compared with state-of-arts multi-label dimensionality methods.

Coronary heart disease (CHD) is a common cardiovascular disease that is extremely harmful to humans, especially middle-aged and old people, with high mortality. Research on the standardization of inquiry information and the design of inquiry model in the diagnosis of CHD will help to realize the standardization and objective of the inquiry information of CHD in TCM diagnosis, and provide methodology reference to the establishment of quantitative diagnosis of CHD, which is of great significance to the promotion of the TCM basis and clinical research of CHD. CHD belongs to the scope of chest heartache in traditional Chinese medicine (TCM); there have been extensive experiences in the diagnosis and treatment of CHD in TCM and the therapeutic effects are fairly satisfying. However, there are few systematic studies of quantitative diagnosis for CHD, especially of the standardization study of inquiry diagnosis for CHD. Some authors investigated the contribution of symptoms to syndromes diagnosis by focusing a complex system on entropy and applying various techniques of multivariate statistics in the construction of diagnostic models in TCM, such as discriminant analysis and regression analysis in the diagnosis of blood stasis syndrome and stroke. Although multivariate statistics has some superiority in the solution of quantitative diagnosis in TCM, the problem on clinical data analysis with high nonlinearity could not be solved by these techniques. Moreover, the complex interaction among different symptoms could not be reflected clearly [10], and the diagnostic rules of TCM could not be revealed comprehensively and widely.

With the introduction of data mining techniques, research workers have applied several nonlinear learning techniques to the research of diagnostic standardization and objectification in TCM, such as k nearest neighbor (kNN), neural networks, Bayesian networks, structure equations, decision tree, genetic algorithm, etc. Most of the algorithms are to solve problems of single syndrome diagnosis, i.e., single label learning. However, in clinical practice, many symptoms are presenting various syndromes. Ref. [11] shows that the main syndromes of CHD are deficiency accompanied with excess, e.g. deficiency of qi syndrome and blood stasis syndrome, deficiency of qi syndrome and turbid phlegm syndrome, deficiency of yang syndrome and turbid phlegm syndrome, blood stasis syndrome and Qi stagnation syndrome, as the predominant combining forms of their syndromes. This is a multi-label learning problem.

Research on the standardization of inquiry information and the design of inquiry model in the diagnosis of CHD will help to realize the standardization and objective of the inquiry information of CHD in TCM diagnosis, and provide methodology reference to the establishment of quantitative diagnosis of CHD, which is of great significance to the promotion of the TCM basis and clinical research of CHD. Some work has been done that applies the multi-label learning to the computer-aided diagnosis of CHD in TCM [12], few efforts tried to apply the multi-label feature selection to the modeling of CHD in TCM.

Based on the inquiry data of CHD in TCM, a hybrid optimization feature selection algorithm, HOML (hybrid optimization based multi-label feature selection) is presented in this paper. HOML combines the relatively strong global optimization ability of simulated annealing algorithm (SA) and genetic algorithm (GA) and the strong local optimization capability of greedy algorithm, and adopts the multi-label classifier to model CHD in TCM.

2 HOML: Hybrid optimization based multi-label feature selection

Genetic algorithm has been used to analyze feature selection for multi-label data [7], but the algorithms only combine the MLNB algorithm, and the genetic algorithm has its limitation in optimization. The performance of feature selection may be further improved if advantages of different optimization techniques are combined together to search for an optimal subset of features. We propose to combine three search algorithms in this paper: mutation-based simulated annealing, genetic algorithm and the greedy algorithm hill-climbing. HOML combines the ability to avoid being trapped in a local minimum of simulated annealing algorithm with a very high rate of convergence of the crossover operator of genetic algorithm and the strong local search ability of the greedy algorithm to obtain the optimal feature subset. Some work has shown that a hybrid technique generated better feature subsets than separate search algorithms [13].

Selection is an important aspect of evolutionary computation. It dictates what members of the current population affect the next generation. More fit individuals are generally given a higher chance to participate in the recombination process. The primary concern of all selection schemes is what is known as the loss of diversity. As a result, information encoded in the current population is not transferred into the next generation in its entirety. Loss of diversity has been measured and analyzed for a number of popular selection algorithms [14, 15]. For the problem of loss of diversity, unbiased tournament selection [16] yields better results, and it is used in HOML.

In the feature selection process of HOML, experiment shows better result when using Average Precision than using (hammingloss+rankingloss). And Average Precision is used as the fitness of feature subset. That is to say, in the training process, we adopt the test result Average_Precision, which is obtained by modeling the validation set using multi-label learning techniques such as ML-KNN [17], BP-MLL [18], Rank-SVM [19] and MLNB-BASIC [9], as the fitness function to evaluate the performance of feature subset. More information about the criteria can be found in subsection 3.2.

Hill climbing is a recursive process, as shown in Figure 1. Figure 2 shows the algorithm flow of HOML, which organizes a search in three stages.

```
HC (FN)
Input:
FN: Feature subset
Output:
BF: The optimized feature subset
    while (Th>0)
    HC(BF) {
      if (Th>0) then
        [NF] = CreateNeighbours(BF);      %Change one feature each time and get N neighbors of BF
        [EM] = EvaluateFitness(NF);       %Evaluate the new feature subset
        [ACC] = Replace(FS, NF);          %If EM(i)>E(i), replace FS[i] with NF[i]. ACC represents set
                                          % which contains improved feature subset.
        for i=1:Num(ACC)                  %Make hill-climbing on each improved feature subset
         BF = HC(ACC(i));
        end for;
        UpdateTime (Th);                  %Update time available for HC.
      end if;
    end for;
```

Figure 1 Procedure of Hill climbing.

Stage 1, HOML employs a simulated annealing (SA) to guide the global search in a solution space. As long as the temperature is very high, SA accepts every solution, thus yielding a near random search through the search space. On the other hand, as the temperature becomes close to zero, only improvements are accepted. The SA is run for approximately 50% of the total time available.

Stage 2: HOML employs a GA to perform optimization. The GA population is set at 100. The initial population consists of the best solutions detected by SA. The crossover operator enables the good solutions to exchange information, and the mutation operator in GA introduces new genes into the population and retains genetic diversity. The GA runs for about 30% of total time spent by HOML to find the optimal feature subset solution.

Stage 3: HOML applies a hill-climbing feature selection algorithm. The greedy algorithm performs a local search on the k-best solutions on the k-best (k represents the dimensionality of feature) solutions given by two global optimization algorithms (SA and GA) and selects the best neighbors. The hill-climbing algorithm is run in the remaining execution time.

The HOML algorithm is implemented on the platform of MATLAB, which is downloaded at http://levis.tongji.edu.cn/gzli/code/homl-code.zip.

3 Dataset and experimental settings

3.1 The used data sets

UCI yeast dataset: In order to further evaluate the introduced algorithm, a western biomedical yeast dataset is used in this paper. The dataset contains 2417 genes each being represented by a 103-dimensional feature vector. Of the 103 features there are no discrete attributes. There are 14 possible class labels. The minimum number of labels for each instance is 0, and the maximum number of labels for each instance is 14. The average number of labels for each gene is 4.24±1.57. More detailed descriptions on this dataset are available in [17].

TCM CHD dataset: The dataset of coronary heart disease are desribed and preprocessed in [12]. A total of 555 cases were obtained, among which 265 patients were male, and 290 patients are female. There are 125 symptoms and 15 syndromes in differentiation diagnosis, of which 6 commonly-used patterns are selected in our study, including: z1 Deficiency of heart qi syndrome; z2 Deficiency of heart yang syndrome; z3 Deficiency of heart yin syndrome; z4 Qi stagnation syndrome; z5 Turbid phlegm syndrome and z6 Blood stasis syndrome. Experimental results show that the predication accuracy was the highest on the

```
HOML (X, Y, Tk, Tg, Th)
Input:
X: $N \times D$ feature matrix
Y: $N \times Q$ label matrix
Tk: Run time for Simulated Annealing (SA)
Tg: Run time for Genetic Algorithm (GA)
Th: Run time for Hill Climbing (HC)
Output:
BF: Optimal feature subset
Procedure: FS = InitIndividual();            %Initialize FS with 100 feature subsets
%Simulated annealing
E = EvaluateFitness(FS);                     % E(i)=Average_Precision(FS(i))
Tc = UpdateTime(Tk);                         %Update the time available time for SA
while (Tc>0)
  FM = Mutate(FS, Pm);          %Mutate FS with probability Pm, Pm = 0.5−0.5exp(Tc/$\lambda$), $\lambda$=Tk/log2(0.5)
  for i=1:100
    if (Fitness(FM(i))>= E(i))
      FS(i) = FM(i);                          %Replace FS(i) with FM(i)
    else if (exp(−(E(i)−Fitness(FM(i))<rand)  %Selectively accept the new feature subset
      FS(i) = FM(i);
    end if;
    UpdateTime(Tc);                           %Update time available for SA.
  end for;
end while;
%Genetic algorithm
while (Tg>0)
      [FS,E] = Select(FS);             %Select with unbiased tournament selection
      [FS,E] = Crossover(FS, 0.65);    % Cross with probability 0.65
      [FS,E] = Mutation(FS,0.01);      % Mutate with probability 0.01
      [E] = EvaluateFitness(FS);       % Evaluate the new solutions
      UpdateTime(Tg);                  % Update time available for GA
%Hill climbing
while (Th> 0){
  Order(FS);                           %Sort FS by fitness value descending.
  for i=1:100
    BF = FS(i);
    BF = HC(BF);
  End for;
End while.
```

Figure 2 ~~F~~Procedure of HOML.

set of 52 symptoms [12]. We made experiments based on the set with 52 symptoms, 3 redundant features like "edema" were manually removed before the experiments, and then we got a dataset with 49 symptoms, which may be downloaded at http://levis.tongji.edu.cn/gzli/data/chd-data.zip. The minimum number of labels for each instance is 0, and the maximum number of labels for each instance is 5. The average number of labels of the sample is 2.58. The attributes of the sample are all discrete.

3.2 Experimental settings

In this paper, several state-of-the-art multi-label learning algorithms including ML-KNN [17], BP-MLL [18], Rank-SVM [19] and MLNB-BASIC [9] are adopted by HOML as base classifiers and compared on the dataset of CHD in TCM. We compare HOML with the following benchmark algorithms: simulated annealing (SA) [20], genetic algorithm (GA) [21], sequential floating forward selection (SFFS) [22], sequential floating backward selection (SFBS) [22], multi-label dimensionality reduction via dependence maximization (MDDM) [3] and embedded feature selection for multi-label learning [7]. The target dimensionality d of MDDM is decided by $thr = 99\%$ [3]. The inner product of label matrix Y is set as the kernel function [3].

Parameters of the multi-label classifiers are set as follows: 1) For ML-KNN, the best parameter k and smoothing factor in [17] are used, which are 10 and 1. 2) For BP-MLL, the number of hidden neurons is set to 8 after which its performance does not significantly change. 3) For Rank-SVM, the type is set to linear SVM. 4) For MLNB-BASIC, the smoothing factor is set to 1.

10-fold cross-validation is carried out to compute the fitness value. In each fold, the time simulated annealing (SA), genetic algorithm (GA) and hill-climbing are allocated 5, 3 and 2 h, respectively. SA adopts the same mutation probability and selective acceptance strategy as in HOML when SA is used as independent optimization techniques. GA adopts the same selection operator, crossover probability and mutation probability as they are in HOML when it is used as independent optimization techniques. When simulated annealing (SA) and genetic algorithm (GA) are used as independent optimization techniques, in each fold, they are allocated 10 h, which is the same time duration for SFFS, SFBS, MDDM, and MEFS. In the training process, 2/3 of the training data are taken as training set, 1/3 as validation.

We also made paired t test on the experimental result on the base classifiers to compare the performance of the feature selection/feature reduction algorithms.

3.3 Evaluation metrics

The following multi-label evaluation metrics proposed in [13] are used in this paper: 1) Hamming loss evaluates how many times an instance-label pair is misclassified. 2) One error evaluates how many times the top-ranked label is not in the set of proper labels of the instance. 3) Coverage evaluates how far we need, on the average, to go down the list of labels in order to cover all the proper labels of the instance. 4) Ranking loss evaluates the average fraction of label pairs that are reversely ordered for the instance. 5) Average precision evaluates the average fraction of labels ranked above a particular label $y \in Y$ which actually are in Y.

4 Experimental results and discussions

We compare our proposed algorithm (HOML) with SA [20], GA [21], SFFS [22], SFBS [22], MDDM [3] and MEFS [7]. Those algorithms are evaluated with ML-KNN [17], BP-MLL [184], Rank-SVM [19], and MLNB-BASIC [9] as base classifiers. Hamming lossone-errorcoverageranking loss and average precision [17] are adopted as the evaluation criteria of multi-label learning model.

4.1 Experimental results on yeast dataset

4.1.1 *Experimental results on accuracy of the features selection/reduction methods*

Firstly we made experiment on the yeast dataset. Table 1 and Table 2 show statistical test results for original results and the results after feature selection/reduction, where CRI means performance criteria, FS means feature selection/reduction method, ORI represents original results of the four classifiers, and Average stands for the average value of the four classifiers under the same condition. Best results on each metric are also in bold.

Table 1 and Table 2 show that compared with the original classification results, predication accuracy has been improved after being optimized by SA, GA, SFFS, SFBS, MEFS and HOML. But the predication accuracy of Hamming loss, one-error, coverage, average precision has been decreased after feature reduction method MDDM, and that the corresponding feature selection/feature reduction methods of the optimal value of the five evaluation criteria are all HOML. The optimal values of the five evaluation criteria, which are the corresponding values of HOML, are as follows: hamming loss 0.2176, 0.0429 lower than the original result 0.2605; one-error 0.2173, 0.1312 lower than the original result 0.3458; coverage 6.4748, 1.0900 lower than the original result 7.5648; ranking loss 0.1851, 0.0635 lower than the original result 0.2486; average precision 75.06%, significantly increased by 10.62 than the original result 64.44%.

Table 1 Comparison of HOML and other feature selection/reduction methods on the yeast dataset: Hamming loss, One-error[a)]

FS	CRI				
	Hamming loss ↓				
	ML-KNN	BP-MLL	Rank-SVM	MLNB-BASIC	Average
ORI	0.1981	0.2405	0.3103	0.2934	0.2605
SA	**0.1841**	0.2309	0.2448	0.2395	0.2248
GA	0.1978	0.2389	0.2499	0.2591	0.2364
SFFS	0.2161	0.2347	0.2485	0.2137	0.2282
SFBS	0.1924	0.2205	0.2698	0.2763	0.2397
MDDM	0.2479	0.3029	0.2536	0.3163	0.2802
MEFS	0.1978	0.2247	0.2502	0.2422	0.2287
HOML	0.1961	0.2159	0.2252	0.2332	0.2176
FS	CRI				
	One-error ↓				
	ML-KNN	BP-MLL	Rank-SVM	MLNB-BASIC	AVE
ORI	0.2386	0.2686	0.5398	0.3473	0.3485
SA	0.2206	0.2625	0.3459	0.2747	0.2759
GA	0.2408	0.2697	0.4025	0.3223	0.3088
SFFS	0.2562	0.2521	0.2479	0.2355	0.2479
SFBS	0.2231	0.2479	0.3843	0.3099	0.2913
MDDM	0.2509	0.4373	0.2509	1	0.4847
MEFS	0.2521	0.2364	0.3103	0.2547	0.2633
HOML	0.1660	0.2066	0.2438	0.2531	0.2173

a) For each criteria, "↓" indicates "the smaller the better" while the "↑" indicates "the bigger the better".

For the performance criteria hamming lossone-errorcoverageranking loss, HOML is lower than feature selection methods SA, GA, SFFS, SFBS and feature reduction method MDDM. On the other hand, HOML is higher than that of the compared feature selection/reduction method and has a significant improvement.

4.1.2 *Paired t test*

During the 10-fold validation process of the experiment, we made paired t test on the test dataset of HOML and its compared algorithms. The hamming loss is computed as eq. (1) when making paired t test. The computations of one-error, coverage, ranking loss and average precision are similar to that of hamming loss. The paired t test result is shown in Table 3. FS represents feature selection/reduction methods; C represents performance criteria. If the value of "Better" is HOML, it means HOML has outperformed its compared algorithm, while HOML has no significant difference from its compared algorithm if the corresponding value of "Better" is NaN.

Table 3 shows that HOML outperforms all the other feature selection/reduction methods as for Hamming loss, Coverage and Average precision; for One error, HOML is only frustrated by SFSFS; for Ranking loss, HOML has not shown significant difference from SA and MDDM, but it is superior to GA, SFFS, SFBS and MEFS.

4.1.3 *Statistics of number of reserved features after feature selection/reduction methods*

In order to further analyze the performance of the feature selection/reduction methods, we made statistics on the average number of the reserved features after optimization as shown in Table 4. The number of reserved features for SA is computed as eq. (2), which is computed similarly to other feature selection/reduction methods. FS means feature selection/reduction methods, and NUM represents the number of reserved features after optimization.

Table 2 Comparison of HOML and other feature selection/reduction methods on the yeast dataset: Coverage, Ranking loss and Average precision[a)]

FS	CRI				
	Coverage ↓				
	ML-KNN	BP-MLL	Rank-SVM	MLNB-BASIC	AVE
ORI	6.3598	7.5779	9.0274	7.2942	7.5648
SA	5.8432	6.8807	7.5472	7.0291	6.8250
GA	6.3938	6.9684	7.6721	7.2599	7.0735
SFFS	6.7810	7.2521	6.8802	6.8264	6.9349
SFBS	6.2893	8.9215	8.6363	7.4587	7.8264
MDDM	7.7791	9.6250	7.9800	10.8145	9.0496
MEFS	6.2893	7.2142	6.9433	7.3547	6.9503
HOML	5.8091	6.6120	6.7521	6.7261	**6.4748**
FS	CRI				
	Ranking loss ↓				
	ML-KNN	BP-MLL	Rank-SVM	MLNB-BASIC	AVE
ORI	0.1701	0.2187	0.3584	0.2473	0.2486
SA	0.1446	0.2007	0.2332	0.2101	0.1971
GA	0.1748	0.2071	0.2440	0.2274	0.2133
SFFS	0.1985	0.2130	0.2140	0.1947	0.2050
SFBS	0.1738	0.2837	0.3121	0.2434	0.2532
MDDM	0.2467	0.3469	0.2574	0	0.2127
MEFS	0.1751	0.2247	0.2475	0.2208	0.2170
HOML	0.1385	0.1752	0.2032	0.2237	**0.1851**
FS	CRI				
	Average precision ↑				
	ML-KNN	BP-MLL	Rank-SVM	MLNB-BASIC	AVE
ORI	0.7417	0.6577	0.5559	0.6225	0.6444
SA	0.7609	0.7152	0.6849	0.7159	0.7192
GA	0.7548	0.7063	0.6679	0.6946	0.7059
SFFS	0.7503	0.7097	0.7065	0.7453	0.7279
SFBS	0.7719	0.6835	0.6062	0.6912	0.6882
MDDM	0.6760	0.5898	0.6794	0.4175	0.5906
MEFS	0.7644	0.6822	0.6743	0.6826	0.7008
HOML	0.7984	0.7471	0.7265	0.7307	0.7506

a) For each criteria, "↓" indicates "the smaller the better" while the "↑" indicates "the bigger the better".

Table 3 Paired *t* test of HOML and its compared algorithms on yeast dataset

C	FS					
	Better					
	SA	GA	SFFS	SFBS	MDDM	MEFS
Hamming loss	HOML	HOML	HOML	HOML	HOML	HOML
One error	HOML	HOML	NaN	HOML	HOML	HOML
Coverage	HOML	HOML	HOML	HOML	HOML	HOML
Ranking loss	NaN	HOML	HOML	HOML	NaN	HOML
Average precision	HOML	HOML	HOML	HOML	HOML	HOML

Table 4 shows that among the statistics of the numbers of reserved features after feature selection/reduction, the minimum number is obtained by SFFS, and SFBS has the maximum number. HOML has a relative small number of reserved features, indicating HOML has efficiently improved the classification accuracy with a relatively small subset of features.

Table 4 The average number of reserved features after optimization on the yeast dataset

NUM	FS						
	SA	GA	SFFS	SFBS	MDDM	MEFS	HOML
Number	41	38	7	92	7	78	42

Table 5 Comparison of HOML and other feature selection/reduction methods on the TCM CHD dataset: Hamming loss, One-error, and Coverage[a)]

FS	CRI				
	Hamming Loss ↓				
	ML-KNN	BP-MLL	Rank-SVM	MLNB-BASIC	Average
ORI	0.3148	0.3733	0.3809	0.3118	0.3452
SA	0.3000	0.3492	0.3370	0.2870	0.3183
GA	0.3051	0.3468	0.3235	0.3124	0.3220
SFFS	0.2897	0.3690	0.3421	0.2942	0.3237
SFBS	0.2876	0.3263	0.3845	0.3214	0.3300
MDDM	0.3009	0.3569	0.3012	0.2899	0.3122
MEFS	0.3006	0.3422	0.3265	0.2912	0.3151
HOML	0.1964	0.2577	0.2411	0.2246	0.2295
FS	CRI				
	One-error ↓				
	ML-KNN	BP-MLL	Rank-SVM	MLNB-BASIC	AVE
ORI	0.2536	0.2620	0.3559	0.3028	0.2936
SA	0.2391	0.2603	0.2885	0.2681	0.2640
GA	0.2410	0.3003	0.3609	0.3083	0.3026
SFFS	0.2356	0.2182	0.3746	0.2693	0.2744
SFBS	0.2458	0.2285	0.3638	0.2145	0.2631
MDDM	0.2111	0.2678	0.2111	0.2412	0.2328
MEFS	0.2464	0.2774	0.2484	0.2492	0.2553
HOML	0.2143	0.1200	0.1455	0.1986	0.1696
FS	CRI				
	Coverage ↓				
	ML-KNN	BP-MLL	Rank-SVM	MLNB-BASIC	AVE
ORI	2.8491	3.1236	3.4000	3.6691	3.2604
SA	2.7669	2.8688	2.9745	2.6473	2.8143
GA	2.8284	2.9675	3.0902	2.9545	2.9601
SFFS	2.6334	3.9107	3.2929	2.8188	3.1639
SFBS	2.6537	2.9864	3.2105	3.1764	3.0067
MDDM	2.8556	3.0700	3.8667	3.3944	3.2967
MEFS	2.7190	2.9976	2.2863	2.4630	2.6164
HOML	2.3750	2.5214	2.5179	2.1421	2.3891

a) For each criteria, "↓" indicates "the smaller the better" while the "↑" indicates "the bigger the better".

4.2 Experimental results on TCM CHD dataset

4.2.1 *Experimental results on the accuracy of the features selection/reduction methods*

After the TCM CHD dataset is preprocessed, we made experiment on it and the experimental results are shown in Table 5 and Table 6. FS represents feature selection/reduction method, CRI represents performance criteria, and ORI represents the original classification results. Average means the average value of the four classifiers under the same condition.

Table 5 and Table 6 show that compared with the original classification results, predication accuracy has been improved after feature selection/feature reduction for all the five evaluation criteria: hamming

Table 6 Comparison of HOML and other feature selection/reduction methods on the TCM CHD dataset: Ranking loss and Average precision$^{a)}$

FS	CRI				
	Ranking Loss ↓				
	ML-KNN	BP-MLL	Rank-SVM	MLNB-BASIC	AVE
ORI	0.2271	0.2728	0.3724	0.2209	0.2733
SA	0.2139	0.2365	0.2623	0.2072	0.2300
GA	0.2236	0.2627	0.3028	0.2399	0.2572
SFFS	0.1957	0.1786	0.3251	0.2294	0.2322
SFBS	0.1876	0.2402	0.3368	0.2018	0.2416
MDDM	0.2178	0.2566	0.2207	0.2124	0.2268
MEFS	0.2063	0.2397	0.3294	0.1875	0.2407
HOML	0.1193	0.1536	0.2672	0.1642	0.1760
FS	CRI				
	Average precision ↑				
	ML-KNN	BP-MLL	Rank-SVM	MLNB-BASIC	AVE
ORI	0.7754	0.7651	0.6985	0.5194	68.96%
SA	0.7940	0.7727	0.7583	0.7994	78.11%
GA	0.8055	0.7960	0.7289	0.7418	76.80%
SFFS	0.8027	0.7842	0.7254	0.7890	77.53%
SFBS	0.8146	0.7882	0.7235	0.7087	75.87%
MDDM	0.7856	0.7529	0.7842	0.7746	77.43%
MEFS	0.7933	0.7318	0.7456	0.8231	77.35%
HOML	0.8819	0.8533	0.8604	0.7443	83.50%

a) For each criteria, "↓" indicates "the smaller the better" while the "↑" indicates "the bigger the better".

loss, one-error, coverage, ranking loss, average precision. On the other hand, we can see that the corresponding feature selection/feature reduction methods of the optimal value of the five evaluation criteria are all HOML. The optimal values of the five evaluation criteria, which are the corresponding values of HOML, are as follows: hamming loss 0.2295, 0.1157 lower than the original result 0.3452; one-error 0.1696, 0.1240 lower than the original result 0.2936; coverage 2.3891, 0.8713 lower than the original result 3.2604; ranking loss 0.1760, 0.0913 lower than the original result 0.2733; average precision 83.50%, significantly increased by 14.54% than the original result 68.96%.

HOML outperforms all the other six feature selection/feature reduction methods (SA, GA, SFFS, SFBS, MDDM, and MEFS): HOML is significantly lower than the six feature selection/reduction methods in terms of hamming loss, one-errorcoverage and ranking loss, and has significantly outperformed the six methods in terms of average precision: SA by 5.44%, SFFS by 5.52%, SFBS by 7.63%, MDDM by 6.07%, and MEFS by 6.15%

Excluding the average value, the separate corresponding classifiers of the optimal values of the five evaluation criteria are as follows: the corresponding classifier of the optimal value of hamming loss (0.1964) is ML-KNN; the corresponding classifier of the optimal value of one-error (0.1200) is ML-KNN; the corresponding classifier of the optimal value of coverage (2.1421) is MLNB-BASIC; the corresponding classifier of the optimal value of ranking loss (0.1193) is ML-KNN; the corresponding classifier of the optimal value of average precision (88.19%) is ML-KNN.

4.2.2 *Paired t test*

We also made paired t test on HOML and its comparison algorithms on the experimental results on the TCM CHD dataset. Table 7 shows the detailed information. The hamming loss is computed as eq. (6) when making paired t test. The computations of one-error, coverage, ranking loss and average precision are similar to hamming loss. FS represents feature selection/reduction methods; C represents performance criteria. If the value of "Better" is HOML, it means HOML has outperformed its comparison

Table 7 Paired t test of HOML and its compared algorithms on the TCM CHD dataset

C	FS					
	Better					
	SA	GA	SFFS	SFBS	MDDM	MEFS
Hamming loss	HOML	HOML	HOML	HOML	HOML	HOML
One error	HOML	HOML	HOML	HOML	NaN	HOML
Coverage	HOML	HOML	HOML	HOML	HOML	NaN
Ranking loss	HOML	HOML	HOML	HOML	NaN	HOML
Average precision	HOML	HOML	HOML	HOML	HOML	HOML

Table 8 The average number of reserved features after optimization on the TCM CHD dataset

NUM	FS						
	SA	GA	SFFS	SFBS	MDDM	MEFS	HOML
Number	19	23	4	38	5	32	20

algorithm, while HOML has no significant difference with its comparison algorithm if the corresponding value of “Better” is NaN.

Table 7 shows that HOML outperforms all its comparison feature selection/reduction methods for performance criteria hamming loss and average precision. For one-error and ranking loss, HOML has not shown obvious advantage over MDDM; for Coverage, HOML has not shown significant difference from MEFS, but is superior to SA, GA, SFFS, SFBS and MDDM.

4.2.3 *Statistics of number of reserved features after feature selection/reduction methods*

In order to further analyze the performance of the feature selection/reduction methods, we also made statistics on the average numbers of the reserved features on the TCM CHD dataset after optimization (Table 8). The number of reserved features for SA is computed as eq. (7), which is computed similarly to other feature selection/reduction methods. FS means feature selection/reduction methods, and NUM represents the number of reserved features after optimization.

Table 8 shows that among the numbers of reserved features after using feature selection/reduction methods, SFFS has the minimum reserved number (4), while the SFBS has the maximum reserved number (38). HOML has a relative small number of reserved features, which indicates that HOML has efficiently improved the classification accuracy with a relatively small subset of features.

4.2.4 *Results on different syndromes*

In order to further analyze the experiment results, the predication accuracies of six syndromes (z1: qi deficiency syndrome; z2: yang deficiency syndrome; z3: yin deficiency syndrome; z4: qi depression; z5: intermingled phlegm syndrome and z6: blood stasis syndrome) were separately tested on the optimal feature subset selected by its corresponding feature selection method and classifier: HOML-ML-KNN. Figure 3 shows predication accuracy of average precision of the six syndromes after feature selection by HOML. X axis represents syndrome label, while the Y axis represents predication accuracy. ORI represents the original classification results; R-FS represents results after feature selection by HOML.

Figure 3 displays that predication accuracy has been increased for each of the six syndromes. The predication accuracy of z1 (qi deficiency syndrome), z2 (yang deficiency syndrome), z3 (yang deficiency syndrome), z5 (intermingled phlegm syndrome) and z6(blood stasis syndrome) have been significantly enhanced (respectively 14%, 9%, 10%, 8%, 8%). The predication accuracy is high before feature selection, and has been slightly increased after feature selection (increased by 3%). We can see from predication accuracy for separate syndrome that precision has been greatly improved by HOML.

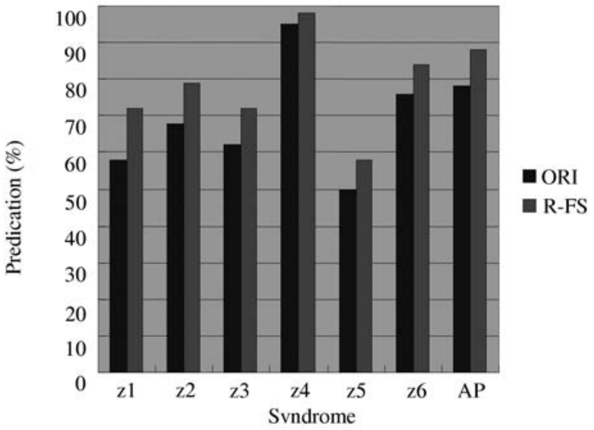

Figure 3 Prediction results of average precision of each syndrome on the optimal feature subset and the original dataset.

5 Conclusions

We have presented a hybrid optimization technique to select the optimal feature subsets in multi-label data, and adopted ML-KNN, BP-MLL, Rank-SVM and MLNB as the multi-label learning classifiers. We compare our algorithm against currently benchmark feature selection/feature reduction techniques on the UCI yeast dataset and TCM CHD dataset, and experimental results suggest our algorithm performs the best. HOML has effectively reduced the data dimension and greatly improved the classification performance. The optimal feature subset of the inquiry symptoms of CHD will be used as a reference in clinical practice.

Future work includes adopting other evaluation criteria or establishing new performance criteria as the fitness function, and combining the feature reduction and feature selection to further improve the performance of modeling.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 60873129, 30901897, 61005006), the Shanghai 3rd Leading Academic Discipline Project (Grant Nos. S30302, B004), and the Open Project Program of National Laboratory of Pattern Recognition in China.

References

1 Tian L, Yan Y J, Zhu J G. Data mining techniques and their application in TCM study (in Chinese). Chinese J Basic Med Trad Chin Med, 2005, 11: 710–712

2 Tsousmakas G, Zhang M L, Zhou Z H. Learning from multi-label data. In: Tutorial at ECML/PKDD'09 Bled, Slovenia 7 2009

3 Zhang Y, Zhou Z H. Multi-label dimensionality reduction via dependence maximization. ACM Trans Knowl Discov Data, in press

4 Yu K, Yu S P, Tresp V. Multi-label informed latent semantic indexing. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2005. 258–265

5 Ji S W, Ye J P. Linear dimensionality reduction for multi-label classification. In: Proceedings of the 21st International Conference on Artificial Intelligence, Pasadena, CA, 2009. 1077–1082

6 Guyon I, Elisseeff A. An introduction to variable and feature selection. J Machine Learn Rese, 2003, 3: 1157–1182

7 Ge L, Li G Z, You M Y. Embedded feature selection for multi-label learning (in Chinese). J Nanjing Univ (Nat Sci), 2009, 45: 671–676

8 Moody J, Utans J. Principled architecture selection for neural networks: Application to corporate bond rating prediction. In: Moody J E, Hanson S J, Lippmann P R, eds. Neural Information Processing Systems 4. Morgan Kaufmann Publishers, Inc, 1992. 683–690

9 Zhang M L, Pena J M, Robles V, et al. Feature selection for multi-label naive Bayes classification. Inf Sci, 2009, 179: 3218–3229

10 Li G C, Li C T, Huang LP, et al. An investigation into regularity of syndrome classification for chronic atrophic gastritis based on structural equation model (in Chinese). J Nanjing Univ Trad Chin Med, 2006, 22: 217–220

11 Wang X W, Qu H B, Wang J. A quantitative diagnostic method based on data-mining approach in TCM (in Chinese). J Beijing Univ Trad Chin Med, 2005, 28: 4–7

12 Liu G P, Li G Z, Wang Y L, et al. Modeling of inquiry diagnosis for coronary heart disease in traditional Chinese medicine by using multi-label learning. BMC Complem Altern Med, 2010, 10: 37

13 Gheyas I A, Smith L S. Feature subset selection in large dimensionality domains. Patt Recogn, 2010, 43: 5–13

14 Blickle T, Thiele L. A comparison of selection schemes used in evolutionary algorithms. Evolut Comput, 1996, 4: 361–394

15 Motoki T. Calculating the expected loss of diversity of selection schemes. Evolut Comput, 2002, 10: 397–422

16 Sokolov A, Whitley D. Unbiased tournament selection. In: Proceedings of the 2005 Conference on Genetic and Evolutionary Computation. Washington DC: ACM, 2005. 1131–1138

17 Zhang M L, Zhou Z H. ML-KNN: A lazy learning approach to multi-label learning. Patt Recog, 2007, 40: 2038–2048

18 Zhang M L, Zhou Z H. Multilabel neural networks with applications to functional genomics and text categorization. IEEE Trans Knowl Data Eng, 2006, 1338–1351

19 Elisseeff A, Weston J. A kernel method for multi-labelled classification. Adv Neur Inf Process Syst, 2002, 14: 681–687

20 Ronen M, Jacob Z. Using simulated annealing to optimize feature selection problem in marketing applications. Europ J Oper Res, 2006, 171: 842–858

21 Yang J, Honavar V. Feature subset selection using a genetic algorithm. IEEE Intell Syst Appl, 1998, 13: 44–49

22 Pudil P, Novovicov J, Kittler J, et al. Floating search methods in feature selection. Patt Recog Lett, 1994, 15: 1119–1125