Effective Feature Selection for Preprocessing Step of Classification Using Modified-MCA

Myo Khaing, Nang Saing Moon Kham
University of Computer Studies, Yangon.
myokhaing.ucsy@gmail.com, moonkhamucsy@gmail

Abstract

A novel metric that integrates the correlation and reliability information between each feature and each class obtained from Multiple Correspondence Analysis (MCA) is currently the popular solution to score the features for feature selection. However, it has the disadvantage that p-value which examines the reliability is conventional confidence interval. The main goal of this paper is to introduce a new classifier independent (filter-based) feature selection method, Modified Multiple Correspondence Analysis (Modified-MCA) which is designed to modify MCA, improving the reliability. The efficiency and effectiveness of proposed method is demonstrated through extensive comparisons with MCA and other feature selection methods, using five benchmark datasets provided by WEKA and UCI repository. Naïve Bayes, Decision Tree and JRip are used as the classifiers. The classification results, in terms of classification accuracy and size of feature subspace, show that the proposed Modified-MCA outperforms three other feature selection methods, MCA, Information Gain, and Relief.

1. Introduction

Feature selection is an important step aiming to extract the most important discriminatory information for classification. The motivation for applying feature selection is multifarious. At first place, features can be expensive to acquire. The cost includes measurement acquisition, data preprocessing, transfer and storage, computational reasons, etc. Furthermore, high-dimensional problems need more samples for training to achieve a good generalization capability of a classifier (i.e., the curse of dimensionality). Reduced dimensionality of the feature set can also help to gain better understanding of a given problem in applications.

Instead of altering the original representation of features like those based on projection (e.g., principal component analysis) and compression (e.g., information theory) [1], feature selection eliminates those features with little predictive information, keeps those with better representation of the underlying data structure.

The Multiple Correspondence Analysis (MCA) is currently the popular solution to score the features for feature selection [2]. The main goal of this paper is to design a classifier independent (filter-based) feature selection method, Modified Multiple Correspondence Analysis (Modified-MCA). The proposed approach, Modified-MCA, continues to explore the geometrical representation of MCA and aims to find an effective way to indicate the relation between features and classes. However, the study tries the p-value as smaller as possible by adjusting with the significance level. Therefore, Modified-MCA could be considered as a potentially better approach. This paper is

organized as follows: Related work is introduced in Section 2; the proposed Modified-MCA is presented in Section 3; followed by an analysis of the experimental results in Section 4. Finally, conclusions are given in Section 5.

2. Related Works

There are many approaches to feature selection proposed in the literature, however, all in principle involve two main ingredients:

1. A search strategy which explores the set of all feature subsets in a purposeful manner.
2. A criterion (objective) function which evaluates those feature subsets.

The search strategy is independent of the criterion function used [3]. The best subset of features is found by optimising (usually maximising) the criterion function. The best performance of the selected features can be achieved when both the feature selection and classification stages are optimized together using the same criterion function [4]. The search strategy usually employs feature ranking ([5], [6]) or subset search ([3], [7]) techniques. Both approaches can be premised on either deterministic or randomized principles which guide the search through the feature space.

Feature selection can be either classifier independent ([5], [8]) (i.e., filter approach) or classifier specific ([8], [9]) (i.e., wrapper approach or embedded method), depending on how it is combined with the construction of the classification model.

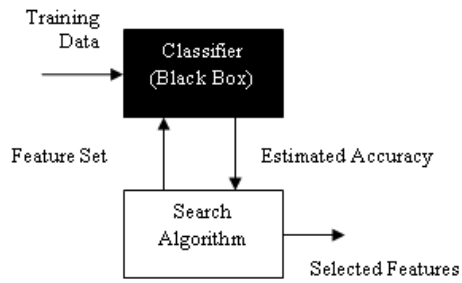


Figure.1 Wrapper Method

Wrappers choose feature subsets with high prediction performance estimated by a specified learning algorithm which acts as a black box, and thus wrappers are often criticized for their massive amounts of computation which are not necessary.

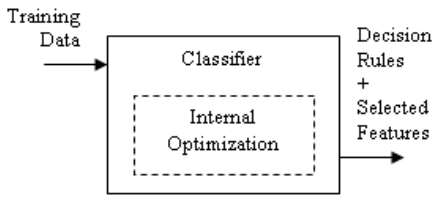


Figure.2 Embedded Method

Similar to wrappers, embedded methods incorporate feature selection into the process of training for a given learning algorithm, and thus they have the advantage of interacting with the classification model, meanwhile being less computationally intensive than wrappers.

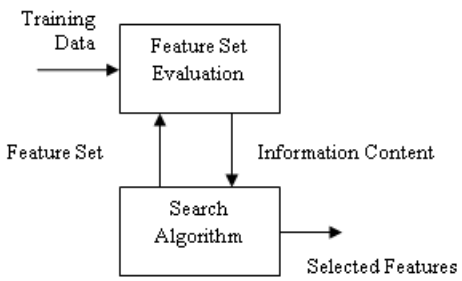


Figure.3 Filter Method

In contrast, filter methods are independent of the classifiers and can be scaled for high-dimensional datasets while remaining computationally efficient. In addition, filtering can be used as a pre-processing step to reduce space dimensionality and overcome the overfitting problem. Therefore, filter methods only need to be executed once, and then different classifiers can be evaluated based on the generated feature subsets [2].

Filter methods can be further divided into two main sub-categories: univeriate and multivariate. The first one is univariate methods which consider each feature with the class separately and ignore the inter-dependence between the features, such as information gain and chi-square measure ([2], [10]).

The second sub-category is the multivariate methods which take features' interdependence into account, for example, Correlation-based feature selection (CFS) and Relief ([11], [12]). They are slower and less-scalable compared to the univariate methods.

3. Modified Multiple Correspondence

Multiple correspondence analysis (MCA) extends the standard Correspondence Analysis (CA) by providing the ability to analyze tables containing some measure of correspondence between the rows and columns with more than two variables.

3.1 Correspondence Analysis (CA)

Standard Correspondence Analysis (CA) is a descriptive/exploratory technique designed to analyze simple two-way contingency tables containing some measure of correspondence between the rows and columns. Multiple Correspondence Analysis (MCA) is an extension of the standard CA [13], and the proposed method Modified-MCA is the modification of MCA.

3.2 Geographical Representation of MCA

MCA constructs an indicator matrix with instances as rows and categories of valuables as columns. Here in order to apply MCA, each feature needs to be first discretized into several intervals or nominal values (called feature-value pairs in the study), and then each feature is combined with the class to form an indicator matrix. Assuming the kth feature has j_k feature-value pairs and the number of classes is m, then the indicator matrix is denoted by Z with size $(n \times (j_k + m))$, where n is the number of instances. Instead of performing on the indicator matrix which is often vary large, MCA analyzes the inner product of this indicator matrix, i.e., Z^TZ, called the Burt Table which is symmetric with size $((j_k + m) \times (j_k + m))$. The grand total of the Burt Table is the number of instances which is n, then $P = Z^TZ /n$ is called the correspondence matrix with each element denoted as p_{ij} .Let r_i and c_j be the row and column masses of P, that is, $r_i = \sum j\ p_{ij}$ and $c_j = \sum i\ p_{ij}$. The center involves calculating the differences $(p_{ij} - r_ic_j)$ between the observed and expected relative frequencies, and normalization involves dividing these differences by $\sqrt{r_ic_j}$, leading to a matrix of standardized residuals $s_{ij} = (p_{ij} - r_ic_j) / \sqrt{r_ic_j}$. The matrix notation of this equation is presented in Equation (1).

$$S = D_r^{-1/2} (P - rc^T) D_c^{-1/2} \quad (1)$$

Where r and c are vectors of row and column masses, and Dr and Dc are diagonal matrices with these masses on the respective diagonals. Through Singular Value Decomposition (SVD), $S = U\Sigma VT$ where Σ is the diagonal matrix with singular values, the columns of U are called left singular vectors, and those of V are called right singular vectors. The connection of the eigenvalue decomposition and SVD can be seen through the transformation in Equation (2).

$$SS^T = U\Sigma V^T V\Sigma U^T = U\Sigma^2 U^T = U\Lambda U^T \quad (2)$$

Here, $\Lambda=\Sigma^2$ is the diagonal matrix of the eigenvalues, which is also called principal inertia. Thus, the summation of each principal inertia is the total inertia which is also the amount that quantifies the total variance of S. The geometrical way to interpret the total inertia is that it is the weighted sum of squares of principal coordinates in the full S-dimensional space, which is equal to the weighted sum of squared distances of the column or row profiles to the average profile. This motivates us to explore the distance between feature-value pairs and classes represented by rows of principal coordinates in the full space. The χ^2 distance between a feature-value pair and a class can be well represented by the Euclidean distance between them in the first two dimensions of their principal coordinates. Thus, a graphical representation, called the symmetric map, can visualize a feature-value pair and a class as two points in the two dimensional map.

As shown in Fig 1, a nominal feature with three feature-value pairs corresponds to three points in the map, namely P_1, P_2, and P_3, respectively. Considering a binary class, it is represented by two points lying in the x-axis, where C_1 is the positive class and C_2 is the negative class. Take P_1 as an example. The angle between P_1 and C_1 is a_{11}, and the distance between them is d_{11}. Similar to standard CA, the meaning of a_{11} and d_{11} in MCA can be interpreted as follows.

Correlation: This is the cosine value of the angle between a feature-value pair and a class in the symmetric map. The symmetric map of the first two dimensions represents the percentage of the variance that the feature-value pair point is explained by the class point. A larger cosine value which is equal to a smaller angle indicates a higher quality of representation [2].

Reliability: As stated before, $\chi 2$ distance could be used to measure the dependence between a feature-value pair point and a class point. Here, a derived value from $\chi 2$ distance called the p-value is used because it is a standard measure of the reliability of a relation, and a smaller p-value indicates a higher level of reliability [2].

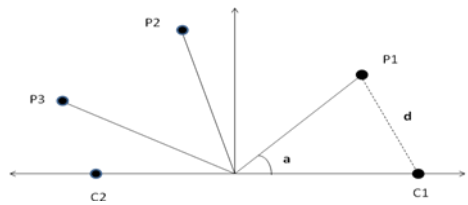


Fig.4 The Symmetric Map of the First Two Dimension

Assume that the null hypothesis H_0 is true. Generally, one rejects the null hypothesis if the p-value is smaller than or equal to the significance level, which means the smaller the p-value, the higher possibility of the correlation between a feature-value pair and a class is true. Here, the conventional significant level is 0.05. It means that a 5% risk of making an incorrect estimate and confidence level of 95%. One never rounds a p-value to zero. Low p-values reported as "$<10^{-9}$", or something similar, indicating that the null hypothesis is 'very, very unlikely to be true', but not 'impossible'. In this paper, the propose M-MCA tries the p-value as smaller as possible by adjusting with the significance level. By this way, standard measure of the reliability can be improved.

P-value can be calculated through the χ^2 Cumulative Distribution Function (CDF) and the degree of freedom is (number of feature-value pairs −1) × (number of classes −1). For example, the χ^2 distance between P_1 and C_1 is d_{11} and their degree of freedom is $(3 - 1) \times (2 - 1)$, and then their p-value is $1-\text{CDF}(d_{11}, 2)$. Therefore, correlation and reliability are from different points of view, and can be integrated together to represent the relation between a feature and a class.

3.3 Modified-MCA Based Feature Selection Model

Modified-MCA continues to explore the geometrical representation of MCA and aims to find an effective way to indicate the relation

between features and classes which contains three stages: Modified-MCA calculation, feature evaluation, and stopping criteria. First, before applying Modified-MCA, each feature would be discretized into multiple feature-value pairs. For each feature, the angles and p-values between each feature-value pair of this feature to the positive and negative classes are calculated, corresponding to correlation and reliability, respectively. If the angle of a feature-value pair with the positive class is less than 90 degrees, it indicates this feature-value pair is more closely related to the positive class than to the negative class, or vice versa. For p-value, since a smaller p-value indicates a higher reliability, (1 - p-value) can be used as the probability of a correlation being true. The p-value is very close to zero but the probability of the correlation being true is very close to zero as well.

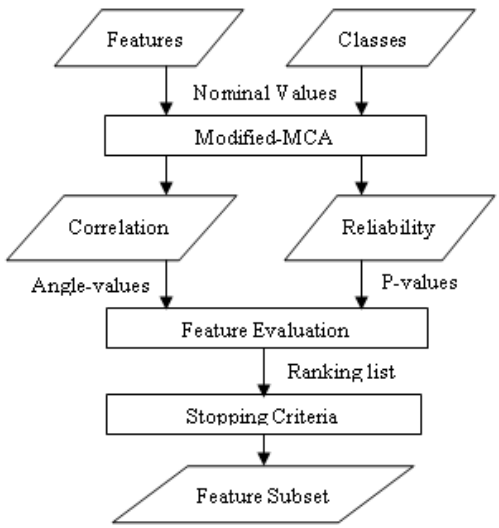


Fig.5 Modified –MCA Based Feature Selection Model

After getting the correlation and reliability information of each feature-value pair, the equations which take the cosine value of an angle and p-value as two parameters are defined (as presented in Equations (3) and (4)) in the feature evaluation stage. Since these two parameters may play different roles in different datasets and both of them lie between [0, 1], different weights can be assigned to these two parameters in order to sum them together as an integrated feature scoring metric. Considering different nominal features contain a different number of feature-value pairs, to avoid being biased to features with more categories like Information Gain does, the final score of a feature should be the summation of the weighted parameters divided by the number of feature-value pairs. Assume there are totally K features. For the k^{th} feature with j_k feature-value pairs, the angles and p-values for the i^{th} feature-value pair are a_{i1} and p_{i1} for the positive class, and a_{i2} and p_{i2} for the negative class, respectively. Then the score of the k^{th} feature can be calculated through Equation (3) or (4).

$$Score\left(k^{th}\ feature\right) = \sum_{1}^{j_k}\left(w_1 \cos a_{i1} + w_2 \max\left(\left(1-p_{i1}\right), p_{i2}\right)\right) / j_k \quad (3)$$

$$Score\left(k^{th}\ feature\right) = \sum_{1}^{j_k}\left(w_1 \cos a_{i2} + w_2 \max\left(\left(1-p_{i2}\right), p_{i1}\right)\right) / j_k \quad (4)$$

If a feature-value pair is closer to the positive class, which means a_{i1} is less than 90 degrees, then equation (3) is applied, where $\max((1-p_{i1}), p_{i2})$ would allow us to take the p-value with both classes into account. This is because that $(1-p_{i1})$ is the probability of the correlation between this feature-value pair and the positive class being true, and p_{i2} is the probability of its correlation with the negative class being false. Larger values of these two probabilities both indicate a higher level of reliability. On the other hand, if a_{i1} is larger than 90 degrees, which means the feature-value pair is closer to the negative class, then $\max((1-p_{i2}), p_{i1})$ will be used instead, that is Equation (4). w_1 and w_2 are the weights assigned to these two parameters. Finally, after getting a score for each feature, a ranked list would be generated according to these scores, and then

different stopping criteria can be adopted to generate a subset of features [2].

4. Experiments and Results

In this section, proposed method is evaluated in terms of speed, number of selected features, and learning accuracy on selected feature subset. Three representative feature selection algorithms, MCA, Information Gain, Relief are chosen in comparison with Modified-MCA. The proposed Modified-MCA is evaluated using five different benchmark datasets from WEKA and UCI repository. The dataset numbers, dataset names, and number of Features in original datasets are shown in Table.1.

Table.1 Datasets Description

No.	Dataset Name	No. of Features	No. of Instances
1	Diabetes	8	768
2	Labor	16	57
3	Ozone	72	2534
4	Soybean	35	683
5	Weather	5	14

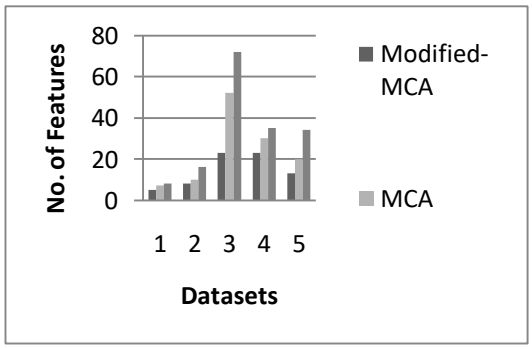
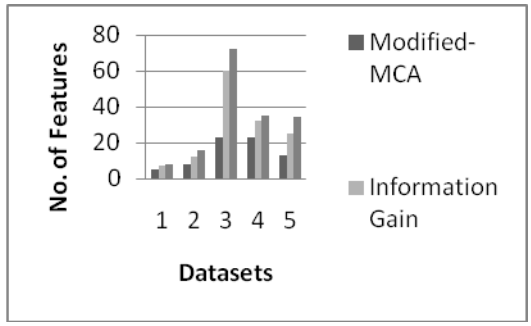


Figure.7 Comparison Results of No. of Features Generated by Modified-MCA and MCA



Figigure.8 Comparison Results of No. of Features Generated by Modified-MCA and Information Gain

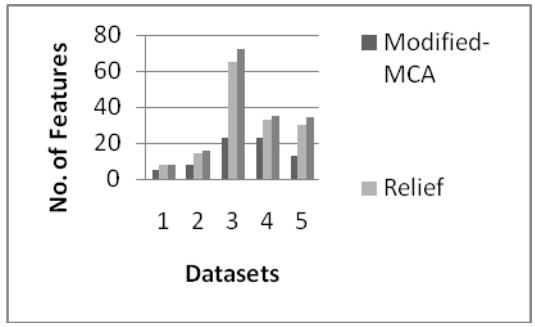


Figure.9 Comparison Results of No. of Features Generated by Modified-MCA and Relief

Three-fold cross validation is first applied to the whole dataset of each concept, which randomly splits the data into three sets with an approximately equal number of data instances and an equal P/N ratio. Then each fold uses two of three sets as the training data set and the remaining one as the testing data set. The final result is the average of these three folds. The proposed method, Modified-MCA, only takes nominal features. In order to get nominal features, discretization on the training dataset needs to be conducted, and then the same intervals are used to discretize the testing dataset. The discretization methods chosen would affect the final classification result. The discretization method applied in this research is the standard discretization method embedded in WEKA which is minimum description length ([14],

[15]). And then, all feature selection algorithms are performed on the discretized training dataset.

In Fig. 7, 8 and 9, the comparison results of number of features generated by Modified-MCA, MCA, Information Gain, and Relief, are shown. It can be significantly seen that the proposed Modified-MCA can generate lesser number of meaningful features than other three feature selection methods, while Relief performs the worst.

After applying, these five sets of data, one for each feature selection method, are run under three classifiers, namely Decision Tree (DT), Rule based JRip (JRip), Native Bayes (NB). Each time, the precision, recall, F-Measure and running time for each classifier based on a particular subset of the features can be obtained.

Table.2 Average Performance of Modified-MCA Based Feature Selection

Dataset	Modified-MCA			
	Precision	Recall	F-Measure	Running Time (sec)
1	0.754	0.756	0.754	0.036
2	0.860	0.859	0.859	0.016
3	0.917	0.869	0.880	0.490
4	0.893	0.871	0.870	0.213
5	0.510	0.667	0.566	0.010
Avg	0.787	0.804	0.786	0.153

Table.3 Average Performance of MCA Based Feature Selection

Dataset	MCA			
	Precision	Recall	F-Measure	Running Time (sec)
1	0.750	0.743	0.746	0.045
2	0.850	0.850	0.850	0.025
3	0.901	0.834	0.866	0.602
4	0.850	0.855	0.852	0.324
5	0.501	0.647	0.564	0.030
Avg	0.770	0.785	0.775	0.205

Table.4 Average Performance of Information Gain Feature Selection

Dataset	Information Gain			
	Precision	Recall	F-Measure	Running Time (sec)
1	0.733	0.737	0.734	0.13
2	0.843	0.841	0.838	0.006
3	0.915	0.846	0.846	2.716
4	0.911	0.889	0.889	0.356
5	0.542	0.690	0.598	0.001
Avg	0.788	0.8006	0.781	0.6418

Table.5 Average Performance of Relief Feature Selection

Dataset	Relief			
	Precision	Recall	F-Measure	Running Time (sec)
1	0.736	0.741	0.737	0.07
2	0.843	0.841	0.838	0.006
3	0.916	0.845	0.864	1.63
4	0.903	0.882	0.901	0.346
5	0.542	0.690	0.598	0.001
Avg	0.786	0.798	0.787	0.416

In Table.2 to 5, the evaluations are discussed by means of average Recall, average Precision, average F-measure and average running time over three classifiers rather than that of only one classifier to be more accurate. Based on the classification results, we can see significantly that the proposed Modified-MCA do better than MCA and other feature selection methods. Although the average F-measure of proposed method is nearly equal to that of Relief, the running time taken to build the classification model is significantly less than that of Relief, 0.153 seconds and 0.416 seconds respectively. The difference is 0.263 seconds. Therefore, the proposed method does better than others feature selection methods.

5. Conclusion

In this study, a new feature subset selection algorithm for classification task, Modified-MCA, was developed. The angles from the proposed method have been used as an indicator of correlation between features and classes, and also an indicator of the contribution of the features. The p-values is taken as a measure of reliability of the relation between features and classes. A ranking list of features can be generated according to the scores and then a features subset can be selected. Based on the results of that experiment, the performance of Modified-MCA is evaluated by several measures such as precision, recall and F-measure. Five different datasets are used to evaluate the proposed method. The results are compared to simple MCA, Information Gain and Relief. The results assure that proposed Modified-MCA makes better results than MCA and other feature selection methods over three popular classifiers.

References

[1] Y. Saeys, I. Inza, and P. Larranaga, A review of feature selection techniques in bioinformatics, *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[2] Qiusha Zhu, Lin Lin, Mei-Ling Shyu, Shu-Ching Chen, *Feature Selection Using Correlation and Reliability Based Scoring Metric for Video Semantic Detection*, 2010.

[3] P. A. Devijver and J. Kittler. *Pattern Recognition: A Statistical Approach*. Prentice Hall, 1982.

[4] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1996.

[5] K. Kirra and L. A. Rendell. The feature selection problem: Traditional methods and a new algorithm. In *Proceedings of the 10th National Conference on Artificial Intelligence*, pages 129–134, San Jose, CA, September 1992. MIT Press, Cambridge, MA.

[6] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, March 2003.

[7] P. Pudil, J. Novovi˘cov´a, and J. Kittler. *Floating search methods in feature selection*. Pattern Recognition Letters, 15:1119–1125, November 1994.

[8] A. Jain and D. Zongker. Feature selection: Evaluation, application and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(2):153–158, February 1997.

[9] R. Kohavi and G. H. John. Wrappers for feature subset selection. Artificial Intelligence, 97(1-2):273–324, 1997.

[10] C. Lee and G. G. Lee, Information gain and divergence-based feature selection for machine learning-based text categorization, *Information Processing and Management*, vol. 42, no. 1, pp. 155–165, 2006.

[11] J. Hua, W. D. Tembe, and E. R. Dougherty, Performance of feature-selection methods in the classification of high-dimension data, *Pattern Recognition*, vol. 42, no. 3, pp. 409–424, 2009.

[12] M. A. Hall, Correlation-based feature selection for discrete and numeric class machine learning, in *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, pp. 359–366.

[13] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen, "Correlation-based video semantic concept detection using multiple correspondence analysis," in *Proceedings of the 10th IEEE International Symposium on Multimedia*, 2008, pp. 316–321.

[14] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, 1993, pp. 1022–1027.

[15] I. Kononenko, "On biases in estimating multi-valued attributes," in *Proceedings of the 14th international joint conference on Artificial intelligence*, 1995, pp. 1034–1040.