Modified-MCA Based Feature Selection Model for Classification

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

Abstract

A central problem in machine learning is identifying a representative set of features from which to construct a classification model for a particular task. A good feature set that contains highly correlated features with the class not only improves the efficiency of the classification algorithms but also improve the classification accuracy. Modified-Multiple Correspondence Analysis (M-MCA or MCA with Geometrical Representation) explores the correlation between different features and classes to score the features for feature selection. The dependence between a feature and a class is measured by a derived value from χ^2 distance called the p-value. It is a standard measure of the reliability of a relation and is examined by p-value. The smaller the p-value, the higher the possibility of the correlation between a feature and a class is true. In this paper, the conventional confidence interval of Multiple Correspondence Analysis (MCA) is modified to get smaller p-value and be more reliable. To evaluate the performance of proposed Modified-MCA, experiments are carried out on benchmark datasets identified and provided by WEKA and UCI repository. In the experiments, Naïve Bayes, Decision Table and JRip are used as the classifiers. The proposed Modified-MCA demonstrates promising results and performs better than well-known feature selection, MCA. The results show that the proposed method outperforms in terms of classification accuracy and reduces the size of feature subspace significantly.

Key-Words: - Feature Selection, Correlation, Reliability, P-value, Confidence Interval.

1. Introduction

Feature subset selection is the process of identifying and removing as much irrelevant and redundant information as possible. This reduces the dimensionality of the data and may allow learning algorithms to operate faster and more effectively. In some cases, accuracy on future classification can be improved; in others, the result is a more compact, easily interpreted representation of the target concept.

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.

In recent years, different areas have adopted the feature selection technique to pre-process the data in order to improve model performance. In general data mining and pattern recognition domains, [2] introduced a criterion function of mutual information and proposed a mutual information based feature selection method which could generate a subset of features without taking class labels into account.

In this paper, the proposed approach, Modified-Multiple Correspondence Analysis (M-MCA), continues to explore the geometrical representation of Multiple Correspondence Analysis (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 M-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 Work

If, however, the data is suitable for machine learning, then the task of discovering regularities can be made easier and less time consuming by removing features of the data that are irrelevant or redundant with respect to the task to be learned. This process is called feature selection. The benefits of feature selection for learning can include a reduction in the amount of data needed to achieve learning, improved predictive accuracy, learned knowledge that is more compact and easily understood, and reduced execution time [8].

Depending on how it is combined with the construction of the classification model, feature selection can be further divided into three categories: wrapper methods, embedded methods, and filter methods. 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. 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. 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 [3].

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 [9][3]. The second sub-category is the multivariate methods which take features' interdependence into account, for example, Correlation-based feature selection (CFS) and Relief [10][11]. They are slower and less-scalable compared to the univariate methods.

According to the form of the outputs, the feature selection methods can also be categorized into ranker and non-ranker. A non-ranker method provides a subset of features automatically without giving an order of the selected features. On the other hand, a ranker method provides a ranked list by scoring the features based on a certain metric, to which information gain, chi-square measure, and relief belong [3].

The different stopping criteria can be applied in order to get a subset from it. Most commonly used criteria include forward selection, backward elimination, bi-directional search, setting a threshold, genetic search, etc.

3. Modified Multiple Correspondence Analysis

In this section, Geometrical Representation of MCA and Modified –MCA Based Feature Selection Model are discussed.

3.1. Geometrical 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 $P = Z^T Z /n$ 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 $(p_{ij} - r_i c_j)$ between the observed and expected relative frequencies, and normalization involves dividing these differences by $\sqrt{}\ r_i c_j$, leading to a matrix of standardized residuals $s_{ij} = (p_{ij} - r_i c_j) / \sqrt{}\ r_i c_j$. 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 D_r and D_c are diagonal matrices with these masses on the respective diagonals. Through Singular Value Decomposition (SVD), $S = U\Sigma V^T$ 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 [3].

Reliability: As stated before, χ^2 distance could be used to measure the dependence between a feature-value pair point and a class point. Here, a derived value from χ^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 [3].

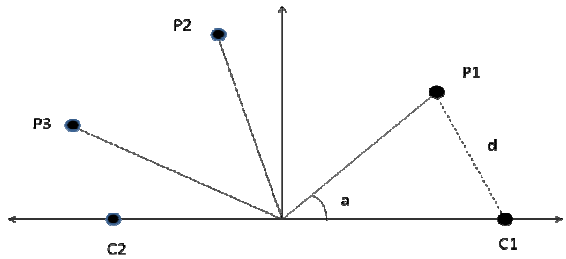


Fig 1. The symmetric map of the first two dimension

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) × (2 − 1), and then their p-value is 1−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.

3.2. Modified –MCA Based Feature Selection Model

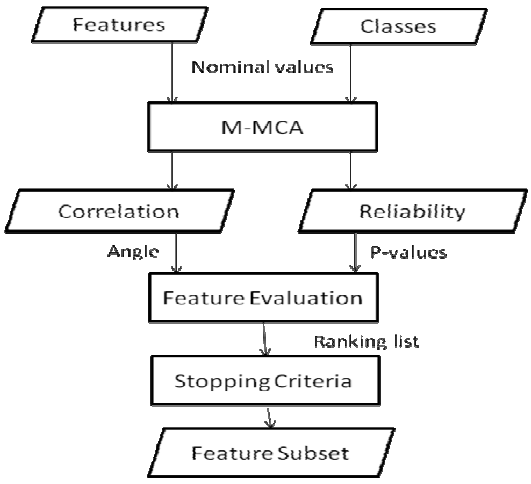


Fig 2. Modified –MCA based feature Selection model

In Fig 2, 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: M-MCA calculation, feature evaluation, and stopping criteria. First, before applying M-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.

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 ai1 and pi1 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. The pseudo code of integrating the angle value and p-value as a feature scoring metric [7] is shown in Fig 3.

Calculating Score

```
1 for k =1 to K
2   for i =1 to jk
3     if cos ai1 > 0
4         sumk+= w1 × cos a i1
5         if count i1 > 0.01 AND count i2 > 0.01
6             sumk+= w2 × max((1 − p i1),p i2)
7     elseif cos a i1 < 0
8         sumk+= w1 × cos a i2
9         if count i1 > 0.01 AND count i2 > 0.01
10            sumk+= w2 × max((1 − p i2),p i1)
11    else
12          sumk+=0
13  end
14     scorek = sumk/jk
15end
```

Fig 3. Calculation score algorithm

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 [3].

4. Experiments and Results

The proposed M-MCA is evaluated using seven different benchmark datasets from WEKA and UCI repository. The dataset numbers, dataset names, and No. of Features in original datasets are shown in Table 1.

Table 1. Datasets description

No.	Dataset Name	No. of Features
1	Diabetes	8
2	Labor	16
3	Ozone	72
4	Soybean	35
5	Weather	5
6	Ionosphere	34
7	Contact-lenses	5

In order to get nominal features, discretization on the training data set needs to be conducted. Next, MCA and M-MCA are performed on the discretized training data set. After applying, these seven sets of data, one for each feature selection method, are run under three classifiers, namely Naïve Bayes (NB), Decision Table (DT), Rule based JRip (JRip). The stopping criterion used for the ranker methods is backward elimination. Each time, the precision, recall and F-Measure of each classifier based on a particular subset of the features can be obtained.

In Table.2 and 3, the evaluation is discussed by means of average Recall, average Precision and average F-measure over three classifiers rather than from individuals.

Precision, Recall and F-measure

In statistics, the F1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision *p* and the recall *r* of the test to compute the score: *p* is the number of correct results divided by the number of all returned results **and** *r* is the number of correct results divided by the number of results that should have been returned. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0.

The traditional F-measure or balanced F-score (F1 score) is the harmonic mean of precision and recall:

$$precision = \frac{|X \bigcap Y|}{|Y|} \quad (5)$$

$$recall = \frac{|X \bigcap Y|}{|X|} \quad (6)$$

$$F = 2.\frac{precision.recall}{precision + recall} \quad (7)$$

Where, X is relevant features, Y is retrieved features, and |X| and |Y| mean the number of features in set X and Y.

Based on the classification results, we can see significantly that the proposed M-MCA perform better than MCA and other feature selection methods, since MCA is better than others [3].

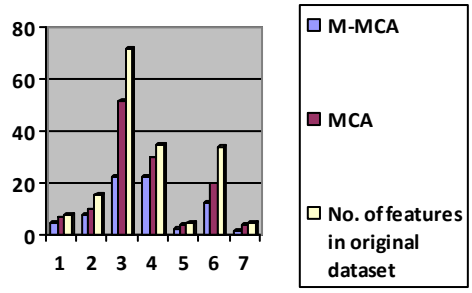
Table 2. Average performance of Modified-MCA based feature selection

Dataset	Modified-MCA		
	Precision	Recall	F-Measure
1	0.54	0.16	0.24
2	0.85	0.86	0.85
3	0.92	0.89	0.90
4	0.41	0.64	0.50
5	0.51	0.67	0.59
6	0.92	0.91	0.91
7	0.56	0.68	0.61
Avg	0.67	0.69	0.66

Table 3. Average Performance of MCA based feature selection

Dataset	MCA		
	Precision	Recall	F-Measure
1	0.49	0.11	0.18
2	0.80	0.80	0.80
3	0.90	0.82	0.86
4	0.40	0.62	0.48
5	0.50	0.62	0.55
6	0.86	0.84	0.85
7	0.50	0.61	0.55
Avg	0.64	0.63	0.63

According to Table 2 and 3, it can be seen significantly M-MCA produces better results than MCA not only on individual dataset but also on overall average, by means of precision, recall, and F-measure.



Note: Rows are no. of features and columns are datasets

Fig 4. Comparison of number of features

In Fig 4, the comparison of number of features generated by M-MCA and MCA are shown, comparing with the number of features in original datasets. In the original Diabetes dataset, there are 8 features. M-MCA can reduce it to 5 features, while MCA can reduce to 7. There are 16 features in Labor dataset. M-MCA and MCA reduce to 8 and 10 features respectively. While MCA reduces 72 features of Ozone dataset to 52, M-MCA can significantly reduce to 23 features. For Soybean and Weather datasets, M-MCA can reduce 35 and 5 features of original datasets to 23 and 3, respectively. It is better than MCA can do: 35 to 30 and 5 to 4. In Ionosphere, although MCA reduce 34 features of original dataset to 20, M-MCA can reduce to 13. In Contact-lenses, it can be seen M-MCA reduce 2 more features than MCA do. Therefore, the size of the feature subspace generated by M-MCA outperforms to those by MCA.

5. Conclusion

In this study, a new feature subset selection algorithm for classification task, M-MCA, was developed. Based on the results of that experiment, the performance of M-MCA is evaluated by several measures such precision,

recall and F-measure. Seven different datasets are used to evaluate the proposed method. The results are compared to simple MCA. The average F-measure over three classifiers increased from 0.63 on MCA to 0.66 on M-MCA. The size of feature subspace can also be reduced significantly as shown in Fig. 3. The results assure that proposed M-MCA makes better results than MCA, 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] H. Liu, J. Sun, L. Liu, and H. Zhang, "Feature selection with dynamic mutual information," Pattern Recognition, vol. 42, no. 7, pp. 1330 –1339, 2009.

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

[4] 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.

[5] Lin Lin, Guy Ravitz, Mei-Ling Shyu, Shu-Ching Chen, "Effective feature space reduction with imbalanced data for semantic concept detection," in SUTC '08: Proceedings of IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing, 2008, pp. 262–269.

[6] M. J. Greenacre and J. Blasius, Multiple Correspondence Analysis and Related Methods. Chapman and Hall/CRC, 2006.

[7] D. Lindley, "A statistical paradox," Biometrika, vol. 44 (1-2), pp. 187–192, 1957.

[8] Mark A. Hall, Correlation-based Feature Selection for Machine Learning, April 1999.

[9] 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.

[10] 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.

[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.