2011 International Conference on Information Communication and Management
IPCSIT vol.16 (2011) © (2011) IACSIT Press, Singapore

Feature Selection Using Modified-MCA Based Scoring Metric for Classification

Myo Khaing, Nang Saing Moon Kham

University of Computer Studies, Yangon, Myanmar.
myokhaing.ucsy@gmail.com, moonkhamucsy@gmail.com

Abstract. Feature subset selection is a technique for reducing the attribute space of a feature set. In other words, it is identifying a subset of features by removing irrelevant or redundant features. A good feature set that contains highly correlated features with the class improves not only the efficiency of the classification algorithms but also the classification accuracy. 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. In this paper, Modified Multiple Correspondence Analysis (M-MCA) is used to improve the reliability. The efficiency and effectiveness of proposed method is demonstrated through extensive comparisons with MCA 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 well-known feature selection methods, MCA, Information Gain, and Relief.

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

1. Introduction

In real-world data, the representation of data often uses too many features, but only a few of them, may be related to the target concept. There may be redundancy, where certain features are correlated so that is not necessary to include all of them in modeling. 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) [11], feature selection eliminates those features with little predictive information, keeps those with better representation of the underlying data structure.

In this paper, 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 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 [7].

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 [10].

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 [1][10]. The second sub-category is the multivariate methods which take features' interdependence into account, for example, Correlation-based feature selection (CFS) and Relief [8][4]. 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 [10].

3. Modified Multiple Correspondence Analysis

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 k^{th} 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 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).

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

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.

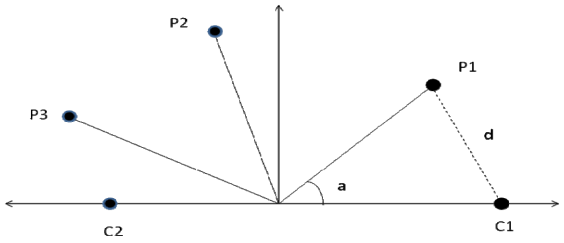

Fig.1: The Symmetric Map of the First Two Dimension

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 [10].

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 [10].

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

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, as shown in Fig 2. 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.

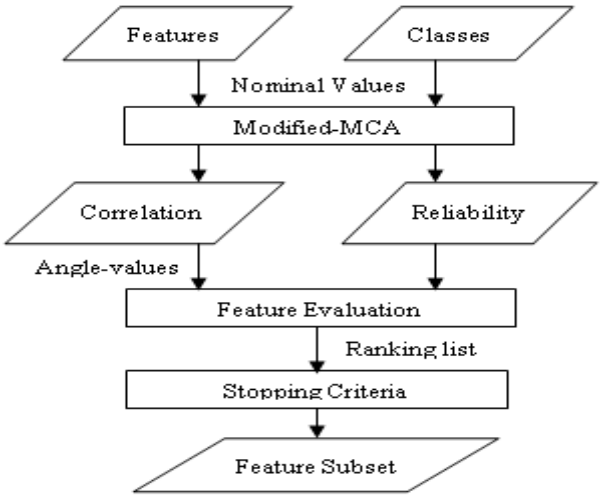

Fig.2: Modified –MCA Based Feature Selection Model

```
Calculating Score
1 for k =1 to K
2   for i =1 to j_k
3     if cos a_i1 > 0
4         sum_k += w_1 × cos a_i1
5         if count_i1 > 0.01 AND count_i2 > 0.01
6             sum_k += w_2 × max((1 − p_i1),p_i2)
7     elseif cos a_i1 < 0
8         sum_k += w_1 × cos a_i2
9         if count_i1 > 0.01 AND count_i2 > 0.01
10            sum_k += w_2 × max((1 − p_i2),p_i1)
11    else
12          sum_k +=0
13  end
14    score_k = sum_k/j_k
15end
```

Fig.3: Feature Score Calculation Algorithm

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).

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

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

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 [2] is shown in Fig.3. 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 [10].

4. Experiments and Results

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

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

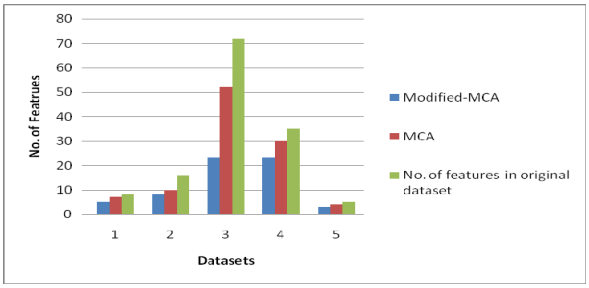

Fig. 4: Number of Features in Original Datasets, Selected by MCA, and Modified-MCA

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

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 M-MCA. The proposed M-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. In Fig. 4, the comparison of number of features generated by Modified-MCA and MCA are shown, comparing with the number of features in original datasets.

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. 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.

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

Based on the classification results, we can see significantly that the proposed M-MCA do better than MCA and other feature selection methods, in terms of average precision, average recall, average F-measure and running time. 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, in terms of numbers, 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, M-MCA, was developed. Based on the results of that experiment, the performance of M-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 M-MCA makes better results than MCA and other feature selection methods over three popular classifiers.

6. References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

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

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

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