FC-MST: Feature Correlation Maximum Spanning Tree for Multimedia Concept Classification

Hsin-Yu Ha, Shu-Ching Chen
School of Computing and Information Sciences
Florida International University, Miami, FL 33199, USA
{hha001,chens}@cs.fiu.edu

Min Chen
Computing and Software Systems, School of STEM
University of Washington Bothell, WA 98011, USA
minchen2@u.washington.edu

***Abstract*—Feature selection is an actively researched topic in varies domains, mainly owing to its ability in greatly reducing feature space and associated computational time. Given the explosive growth of high-dimensional multimedia data, a well-designed feature selection method can be leveraged in classifying multimedia contents into high-level semantic concepts. In this paper we present a multi-phase feature selection method using maximum spanning tree built from feature correlation among multiple modalities (FC-MST). The method aims to first thoroughly explore not only the correlation between features within and across modalities, but also the association of features towards semantic concepts. Secondly, with the correlations, we identify important features and exclude redundant or irrelevant ones. The proposed method is tested on a well-known benchmark multimedia data set called NUS-WIDE and the experimental results show that it outperforms four well-known feature selection methods in all three important measurement metrics.**

I. Introduction

Feature selection is the process of identifying the most appropriate features from the original feature set based on certain evaluation criteria [1]. It has been intensively explored in various research fields, including pattern recognition [2], [3], machine learning [4], [5], data mining [6]–[8] and statistics [9], to name a few. It is usually applied to reduce high-dimensional feature space by selecting only the relevant and important features. Research shows that a well designed feature selection method can not only handle high-dimensional data sets, but also successfully enhance classification performance in coping with imbalanced data where one class has way more data instances than the other class(es) [5], [10]–[13]. Hence, feature selection has been widely applied in applications with imbalanced datasets such as medical decision making using MRI images [14] or EMG signals [15], biomedical studies using microarray gene data sets [16], and text categorization [11], [17], etc.

Generally speaking, feature selection methods can be categorized into three classes, supervised algorithms [18], [19], unsupervised algorithms [20], [21], or semi-supervised algorithms [22], [23]. As supervised algorithms require a set of labeled training data that generally involves expensive human labor, many researchers are increasingly focused on unsupervised or semi-supervised methods in selecting good features. On the other hand, feature selection methods can also be classied into different types of strategies including filter, wrapper, and embedded methods [9]. In filter methods [24], only the general characteristics of training data are considered to evaluate the predefined relevance score of each feature. No learning algorithms or induction algorithms are involved during the process. Therefore, it has a lower computational cost compared to the other two. The wrapper methods [25] work closely with certain classification algorithm whose classification results are used as the evaluation criteria to determine whether a subset of features captures relevant information. The feature subset produces the least classification errors will be selected to build the classification model. Usually, the wrapper methods can outperform the filter methods with regard to classification accuracy. However, the process requires a proper integration of multiple components including a predefined classification algorithm, a good feature relevance criterion, and an efficient searching method to identify feature subset. In addition, it is computationally intensive and may lead to over-fitting problem. Lastly, the embedded methods [26], [27] incorporate learning methods by using objective functions to evaluate feature relevance and select relevant feature subset. Unlike wrapper methods, it doesn't search through the space of all possible feature subsets but identify feature subsets via selected learning strategy. Hence, it is less computationally expensive. In addition, it is also less prone to overfitting compared to wrapper methods.

In this work, we propose a feature selection method called FC-MST to cope with high-dimensionalities and imbalanced problem in multimedia concept detection. The proposed method first applied Multiple Correspondence Analysis (MCA) to project original features into a two-dimensional feature space and obtain feature correlations. Then, a Maximum Spanning Tree is built using the correlations and eliminate irrelevant and redundant features by pruning the tree. The goal is to explore possible feature correlations within and among different modalities and further utilize the correlation to identify the ones that are important and highly relevant to the targeted semantic concepts.

The rest of the paper is organized as follows. We present the overview of the proposed framework and the detail of each component in section 2. In section 3, we explain the design of the experiments and analyze the experimental results. Finally, the paper is concluded in section 4.

IEEE ICSC 2015, February 7-9, 2015, Anaheim, California, USA
978-1-4799-7935-6/15/$31.00 ©2015 IEEE

II. Proposed Framework

For each semantic concept, the proposed FC-MST feature selection method aims to identify a feature subset, containing the important and relevant features from the original multi-modal feature set, to improve the performance of semantic concept classification. It is a three-step supervised method as shown in Figure 1.

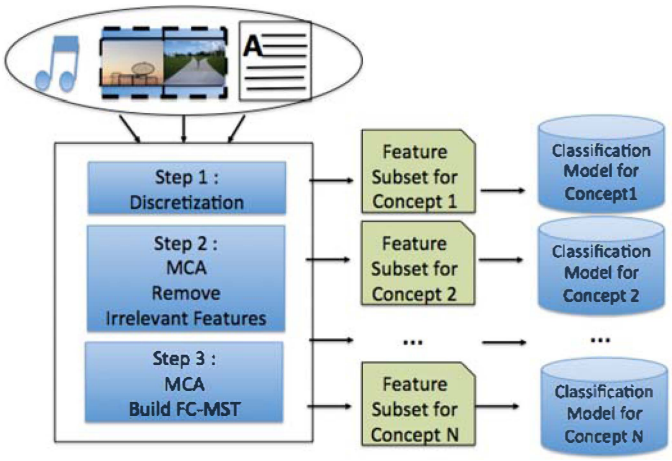

Fig. 1: An overview of the proposed framework

A. Step1: Features Eliminated from Discretization Process

To handle both numeric and nominal features, a supervised method called Minimum Description Length (MDL) [28] is used to discretize each feature into a number of intervals based on its values associated with a target concept. For example, Table I shows 5 instances with M features and two columns at the end indicates the label of positive or negative concept. If an instance has value 1 in the positive concept column, it means the concept can be observed from the instance, and vice versa.

TABLE I: Example of the Original Features

	Feature 1	Feature 2	...	Feature M	Target Concept Positive	Target Concept Negative
Inst. 1	-0.49	1.08	...	-0.45	1	0
Inst. 2	-0.56	-0.85	...	-1.32	0	1
Inst. 3	-0.61	-2.21	...	1.33	1	0
Inst. 4	-0.48	-0.97	...	-1.01	0	1
Inst. 5	-0.53	-1.54	...	0.97	1	0

After discretization, all feature values are grouped into intervals and are denoted as F_j^i where i is the index of feature and j is the index of the interval. For instance, F_3^2 means the third interval of the second feature. Table II shows example discretization results of Table I. As we can see, all instances share the same value in the feature 1 column (i.e., F_1^1). This means feature 1 doesn't have the distinguish ability for the target concept and such features will be removed in the first step of our proposed method as shown in Algorithm 1.

TABLE II: Example of the Discretized Features

	Feature 1	Feature 2	...	Feature M	Target Concept Positive	Target Concept Negative
Inst. 1	F_1^1	F_3^2	...	F_2^M	1	0
Inst. 2	F_1^1	F_2^2	...	F_1^M	0	1
Inst. 3	F_1^1	F_1^2	...	F_3^M	1	0
Inst. 4	F_1^1	F_3^2	...	F_1^M	0	1
Inst. 5	F_1^1	F_1^2	...	F_3^M	1	0

Algorithm 1: Feature eliminated from discretization process

input : The given training data set D with feature set as $TDF = F_1, F_2, ..., F_M$, along with the class label C
output: SF_1: A set of selected features

```
SF_1 ⟵ TDF;
for i ← 1 to M do
    NumFI_i = |MDL(F_i)|;
    /* NumFI_i represents the number of intervals in the i^th feature */
    if NumFI_i = 1 then
        SF_1 ⟵ SF_1 − {F_i};
    end
end
return SF_1
```

B. Step2 : Features Eliminated from MCA

Multiple Correspondence Analysis (MCA) has been applied and proven effective to the research areas ranging from feature selection [29], discretization [30], video semantic concept detection [31]–[38], to data pruning [39]. In this paper, our previous work [29] is integrated as a preprocess step, which has been demonstrated to outperforms other existing feature selection methods, such as information gain (IG), Chi-Square measure (CHI), etc., in terms of classification accuracy.

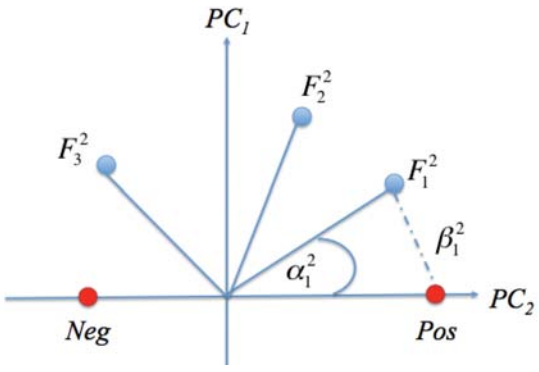

Fig. 2

After applying MCA to a data set as presented in Table II,

Algorithm 2: Features Eliminated from MCA

input : A given training data set D_1 with selected feature set $SF_1 = F_1, F_2, ..., F_L$, along with the class label C

output: SF_2: A set of selected features

$SF_2 \longleftarrow SF_1$;
for $i \leftarrow 1$ **to** L **do**
  $(FIC, FIR) = MCA(D_1)$;
  /* Correlation and reliability of feature interval toward target concept */
  for $j \leftarrow 1$ **to** $NumFI_i$ **do**
    $SumCorrelation+ = FIC_j$;
    $SumReliability+ = FIR_j$;
  end
  $FC_i = (SumCorrelation + SumReliability)/NumFI_i$
end
if $FC_i < TH$ **then**
  $SF_2 \longleftarrow SF_1 - \{F_i\}$;
end
return SF_2

all the intervals of a feature are projected on a two-dimensional space composed by two major principal components, PC_1 and PC_2. Figure 2 depicts three intervals of feature 2 and two red dots which represent positive and negative classes. The relation between an interval of a particular feature and the positive class can be represented by two factors. One is called $Correlation$ α_j^i (e.g., α_1^2), which is the cosine value of the angle between the feature interval F_j^i (e.g., F_1^2) and the positive class. The other is called $Reliability$ β_j^i (e.g., β_1^2), which is the distance between a feature interval F_j^i (e.g., F_1^2) and the positive class. Together these two can be used as a relevance score of a feature interval toward the semantic concept. Zhu et al. [29] go further to obtain the average relevance score per feature to eliminate features whose score is lower than a preset threshold as shown in Algorithm 2. This method is adopted here as a preprocess step to obtain important features for building Maximum Spanning Tree (MST) in step 3.

C. Step3 : Feature Eliminated from FC-MST

1) Building Feature Correlation Adjacency Matrix: In section II-B, MCA is used to capture correlation between feature intervals and the positive target concept as shown in Figure 2. To build the maximum spanning tree, we apply MCA to the remaining features from section II-B to explore correlations between each pair of them. Take Figure 3 as an example, all the intervals of the second feature F^2 and the third feature F^3 are projected onto the two-dimensional symmetric map. The cosine value of each pair of intervals from different features will be generated and the maximum value is selected as the correlation between this pair of features as shown in equation

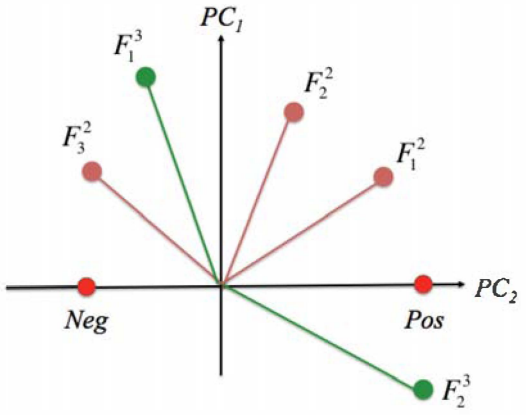

Fig. 3

Algorithm 3: Building Feature Correlation Adjacency Matrix

input : A given training data set D_2 with selected feature set $SF_2 = F_1, F_2, ..., F_L$, along with the class label C

output: Adjacency Matrix AM and the corresponding undirected weighted graph $G(F, E)$

for $i \leftarrow 1$ **to** L **do**
  for $j \leftarrow 1$ **to** L **do**
    $(FIC, FIR)_{ij} = MCA(D_1)$;
    /* Correlation and reliability of feature intervals of one feature toward feature intervals of the other feature */
    if $i = j$ **then**
      $AM(i, j)$ = 0 ;
    else
      $AM(i, j) = Max(FIC, FIR)_{ij}$;
    end
  end
end
return AM

1.

$$FC_{ij} = \begin{cases} \text{argmax}\, Cos(\alpha_{F_m^i F_n^j}), & \text{if } i \neq j \\ 0, & \text{if } i = j \end{cases} \quad (1)$$

Here, i and j are indexed from 1 to L, the total number of the remaining features. The feature correlation between any feature and itself is set to be zero. Therefore, an $L*L$ adjacent matrix can be obtained where each feature is a vertex and the correlation is the edge. Consecutively, an undirected weighted graph $G(F, E)$ is built upon the adjacent matrix where F is the set of remaining features and E indicates the set of feature correlation $\{FC_{ij}\}_{i,j=1}^{L}, i \neq j$.

Algorithm 4: FC-MST
Feature Correlaion Maximum Spanning Tree

input : An undirected weighted graph $G(F, E)$, comprising a set of features SF_2 = F_1, F_2, ..., F_L together with a set of edges which have feature correlation between each feature pairs as the value FC_ij where i and $j \in 1, 2, ...L$, $i < j$. A set of Feature Correlation toward target concept FC_iC where $i \in 1, 2, ...L$
output: SF_3: A set of selected features

1 $SF_3 \longleftarrow \emptyset$; /* Selected features starts with an empty set */
2 $MaxSpanTree = Prim(G)$; /* Applying Prim algorithm on undirected weighted graph G */
3 **for** *each Edge* $E_{ij} \in MaxSpanTree$ **do**
4 | **if** $FC_{ij} < FC_{iC}$ **and** $FC_{ij} < FC_{jC}$ **then**
5 | | $MaxSpanTree \longleftarrow MaxSpanTree - E_{ij}$
6 | **end**
7 **end**
8 $C = BreadFirstSearch(MaxSpanTree)$;
/* Apply BFS algorithm and return a set of components */
9 **for** *Each Component* $C_m \in C$ **do**
10 | $SF_3 \longleftarrow MaxFC(C_m)$
11 **end**
12 return SF_3

2) Building Feature Correlation Maximum Spanning Tree: There are three purposes of building a feature correlation maximum spanning tree as listed below:

- Partition FC-MST into relevant feature clusters which have high intra-cluster correlation and low inter-cluster correlation
- Identify representative features from each feature clusters
- Eliminate redundant and irrelevant features from FC-MST

As shown in Algorithm 4, given the undirected weighted graph from section II-C, a maximum spanning tree is constructed using Prim's method [40] which spans over all the feature vertices based on the correlation values. In brief, the proposed FC-MST is an acyclic subgraph that has the maximum sum of feature correlation weights across all the features nodes. Once the maximum spanning tree is built, the proposed algorithm (see statement 2 in Algorithm 4) loops through all the edges and removes the ones whose weight FC_{ij} is smaller than the correlation of features toward concept, e.g., FC_{iC} and FC_{jC} (see statements 3 to 7 in Algorithm 4). Breadth-first search (BFS) [41] is applied to identify a set of disconnected components (i.e., clusters) $C = C_1, C_2, ..., C_N$ after such edges removal. The feature with the largest correlation toward the target concept in one cluster will be selected as its representative feature. Since every cluster is composed by highly correlated features, all the other features besides the representative one are considered redundant and they are removed from the feature set (see statements 8 to 11 in Algorithm 4). At the end, a subset of representative features is selected to build the classification model for each semantic concept.

III. Experiments

A. Dataset

NUS-WIDE [42], a large-scale image data set containing 269,648 images and the associated tags, is introduced to evaluate the performance of the proposed feature selection method. It has six types of low-level visual features extracted from the images, e.g., color histogram, color correlogram, edge direction histogram, etc., and user tags from flickr website represented as text features. There are 81 high-level semantic concepts, most of them highly imbalanced with the PN ratio (i.e., the number of positive instances vs. negative ones) lower than 1%.

B. Evaluation Criteria

As discussed earlier, a general use of the feature selection method is to identify a subset of representative features that enable classifiers to build better classification models more efficiently. Therefore, we can assess the performance of a feature selection method by evaluating performance of the resulting classification model and efficiency of the classification process. Consequently, the proposed feature selection method is evaluated and compared with other state-of-the-art methods using three criteria.

1) **Classification Model Performance**

TABLE III: Confusion Matrix

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	TruePos	FalseNeg
	Negative	FalsePos	TrueNeg

Confusion matrix (see an example in Table III) is widely used in machine learning and data mining areas to visu-

alize classification results in table-layout fashion. Many performance metrics can be derived from it to analyze the classification results from different perspectives.

- **Precision**
  Based on Table III, precision is calculated as

$$Precision = \frac{TruePos}{(TruePos + FalsePos)} \quad (2)$$

  In other words, precision shows the fraction of retrieved instances that are relevant, where a high precision indicates a lower false positive rate.
- **Average Precision and Mean Average Precision**
  Average precision (AP) and mean average precision (MAP) are two metrics extended from precision, as defined in equation 3 and equation 4, respectively. In brief, **Average Precision** at K is used to evaluate top K ranked results, where $\#(TopR)$ represents the number of instances which are correctly classified as positive instances among top R retrieved instances, $R = 1...K$. A high AP value means more relevant results are ranked earlier than irrelevant ones.

$$AP(K) = \frac{1}{K} \sum_{R=1}^{K} \frac{\#(TopR)}{R} \quad (3)$$

  Mean Average Precision is used to validate ranked results for more than one concepts, where TC is the total number of concepts and $AP_C(K)$ is the average precision at K for concept C.

$$MAP(K) = \frac{\sum_{C=1}^{TC} AP_C(K)}{TC} \quad (4)$$

2) **Feature Reduction Rate** The purpose of feature selection method is to select the most relevant and important features while greatly reducing the feature space. Hence, the proposed method is also evaluated in terms of feature reduction rate, which is calculated in equation 5.

$$FRR = \frac{(OF\# - FS\#)}{OF\#} \quad (5)$$

where $OF\#$ represents the number of original features and $FS\#$ represents the number of remaining features after applying feature selection method.

3) **Efficiency Rate** Lastly, efficiency rate is defined by taking both MAP value and processing time into account as shown in equation 6.

$$ER = \frac{MAP(K)}{ProcessingTime} \quad (6)$$

On one hand, a higher MAP value indicates more positive instances being successfully given higher ranking scores. On the other hand, a reduced feature space leads to shorter processing time. Therefore, given the equation 6, a higher efficiency Rate (ER) represents a better overall performance for a feature selection method.

C. Experimental results

In the experiments, our proposed method is compared with four well-known feature selection methods, e.g., ChiSquare, Filter, InfoGain, and Wrapper. After feature selection on the NUS-WIDE data, Support vector machine (SVM), a constructive learning algorithm, is used to build classification models. SVM is chosen because of its capability in classifying high-dimensional data [43]. Three-fold cross validation scheme is adopted to avoid bias.

First, the experimental result demonstrates the comparison between the proposed method and the other feature selection methods in terms of the MAP values. As shown in Table IV, the proposed method FC-MST achieves the highest MAP values and thus outperforms all other methods in all cases, where K is set to different values in the range of 5 to 200. The proposed method is also the only feature selection method that maintains over 0.7 MAP value across all cases. The trend can also be seen in Figure 4.

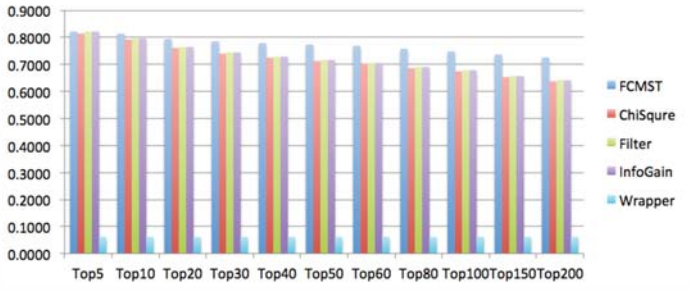

Fig. 4: The MAP values of 81 concepts in NUS-WIDE for different retrieved levels against other feature selection methods

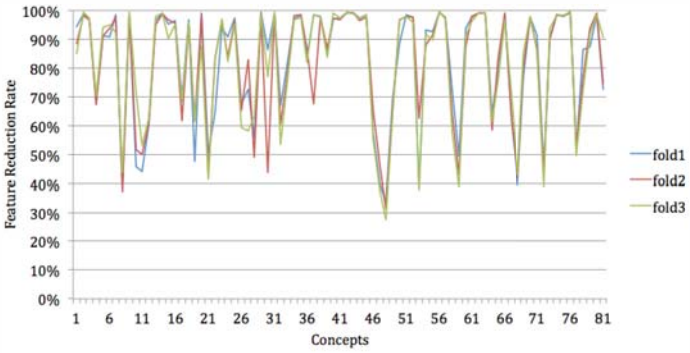

Fig. 5: Feature Reduction Rate (FRR) for NUS-WIDE 81 concepts after applying FC-MST

Secondly, Figure 5 depicts the feature reduction rate (FRR) over all 81 concepts after applying the proposed feature selection method. Among them, we achieved more than 90% FRRs on 40 concepts. The experiment indicates that the proposed method can greatly reduce the original feature space and are especially helpful in dealing with high-dimensional data sets.

TABLE IV: The MAP values of 81 concepts in NUS-WIDE against other feature selection methods

Method	K = 5	K = 10	K = 20	K = 30	K = 40	K = 50	K = 60	K = 80	K = 100	K = 150	K = 200
FC-MST	**0.8217**	**0.8133**	**0.7940**	**0.7854**	**0.7786**	**0.7734**	**0.7688**	**0.7578**	**0.7481**	**0.7361**	**0.7257**
ChiSqure	0.8140	0.7917	0.7604	0.7398	0.7246	0.7125	0.7015	0.6862	0.6744	0.6524	0.6370
Filter	0.8215	0.7961	0.7645	0.7439	0.7287	0.7166	0.7057	0.6903	0.6785	0.6566	0.6412
InfoGain	0.8215	0.7961	0.7645	0.7439	0.7287	0.7166	0.7056	0.6903	0.6785	0.6566	0.6411
Wrapper	0.0617	0.0617	0.0617	0.0617	0.0617	0.0617	0.0617	0.0617	0.0617	0.0617	0.0617

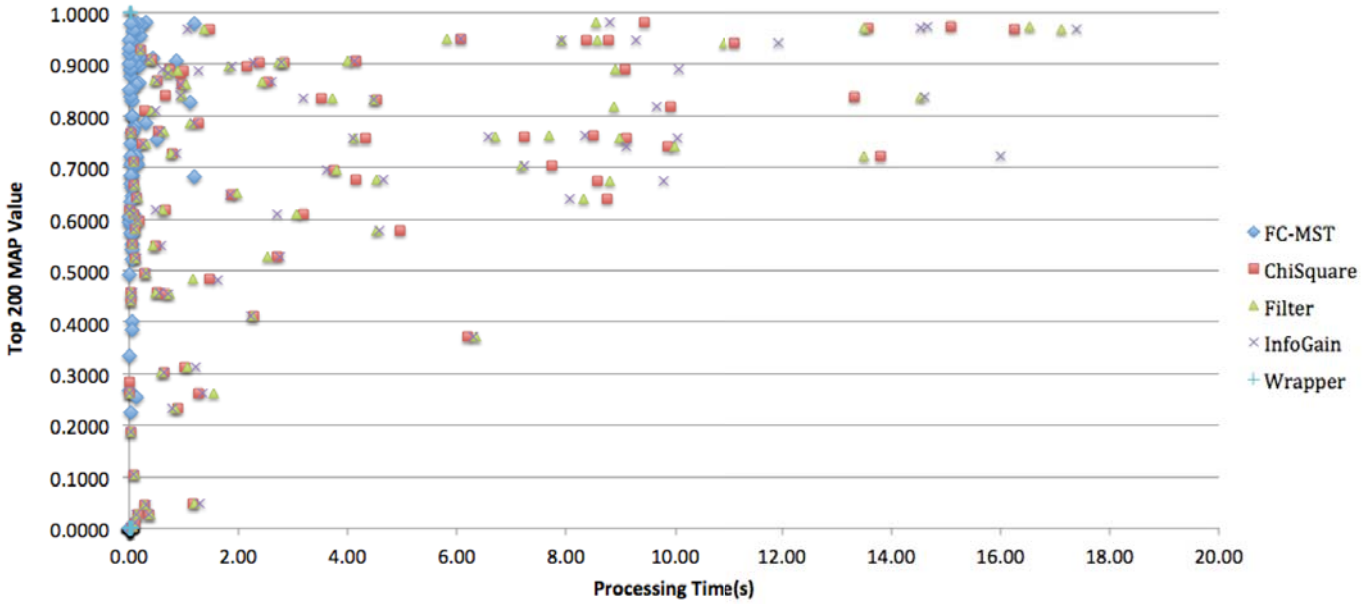

Fig. 6: Top200 Map Value v.s. Processing Time against other feature selection methods

Thirdly, experiment is conducted to validate whether the proposed method is able to reduce the processing time meanwhile producing a compatible classification results against other methods in terms of MAP value. In Figure 6, the results are projected on a two-dimensional chart, where x-axis represents the computation time for the classification process in seconds and y-axis shows the MAP values at K = 200. As shown in Figure 6, the proposed FC-MST method can achieve similar or better MAP value as compared to other methods while using significantly shorter processing time.

Lastly, the efficiency rate is calculated as defined in equation 6 using MAP value at K = 200. In Figure 7, it can be easily observed that FC-MST has the highest efficiency rate across all the 81 concepts except for a few concepts where the wrapper method produces better rates. This is because the wrapper method selects only one feature, its processing time is the shortest. However, as can be seen in Table IV, the wrapper method produces much worse MAP values (always the worst among all methods).

IV. Conclusion

In this paper, we propose a three-steps feature selection method FC-MST. It uses Multiple Correspondence Analysis to explore correlation among features within and across modalities and to capture correlation between feature and targeted semantic concepts. It also allows visual depicts of feature correlation using Maximum Spanning Tree. Consequently, it enhances the classification performance on multimedia data by effectively removing redundant and irrelevant features from high-dimensional data. As shown in the experiments, FC-MST outperforms four other well-known feature selection methods in all three perspectives: MAP, feature reduction rate, and efficient rate. It proves that the proposed method can not only greatly reduce computational cost owing to feature space reduction, but also lead to better classification results.

Acknowledgment

This research was supported in part by the U.S. Department of Homeland Security under grant Award Number 2010-ST-062-000039, the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001, and NSF HRD-0833093.

References

[1] L. Liu, J. Kang, J. Yu, and Z. Wang, "A comparative study on unsupervised feature selection methods for text clustering," in *Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering*. IEEE, 2005, pp. 597–601.

[2] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 2, pp. 153–158, 1997.

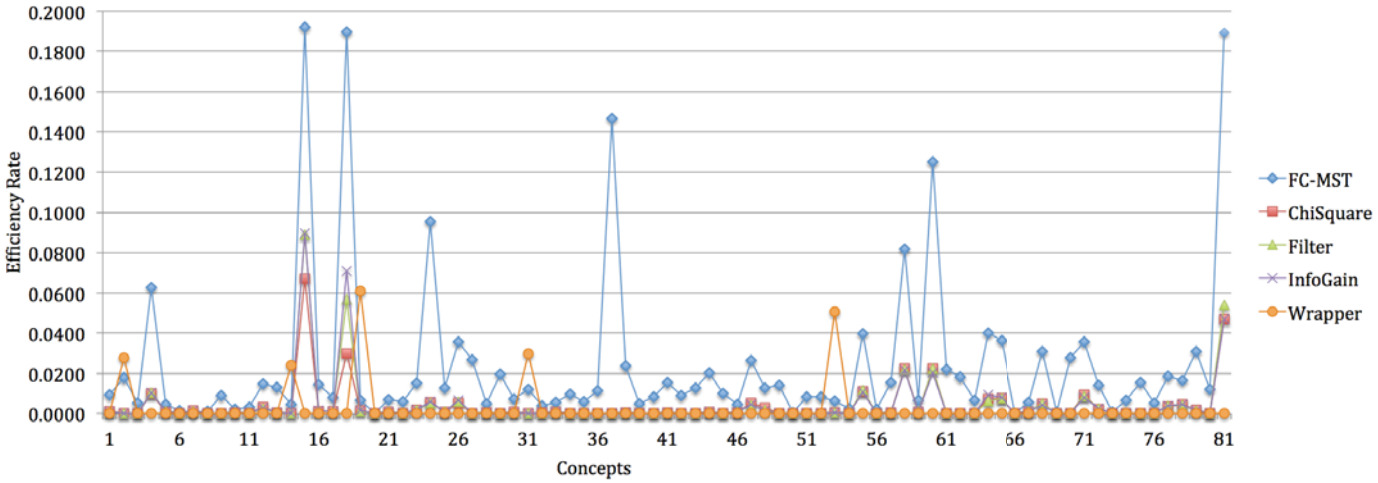

Fig. 7: The efficiency rate of 81 concepts in NUS-WIDE against other feature selection methods

[3] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic press, 1990.

[4] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[5] D. Mladenic and M. Grobelnik, "Feature selection for unbalanced class distribution and naive bayes," in *Proceedings of the 16th International Conference on Machine Learning (ICML*. Citeseer, 1999.

[6] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*. Springer, 1998.

[7] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *ICML*, vol. 3, 2003, pp. 856–863.

[8] M. Chen, S.-C. Chen, and M.-L. Shyu, "Hierarchical temporal association mining for video event detection in video databases," in *2007 IEEE 23rd International Conference On Data Engineering Workshop*. IEEE, 2007, pp. 137–145.

[9] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[10] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *The Journal of machine learning research*, vol. 3, pp. 1289–1305, 2003.

[11] Z. Zheng, X. Wu, and R. Srihari, "Feature selection for text categorization on imbalanced data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 80–89, 2004.

[12] X.-w. Chen and M. Wasikowski, "Fast: a roc-based feature selection metric for small samples and imbalanced data classification problems," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 124–132.

[13] S.-C. Chen, R. L. Kashyap, and A. Ghafoor, *Semantic Models for Multimedia Database Searching and Browsing*. Springer, 2000, vol. 21.

[14] N. Zhang, S. Ruan, S. Lebonvallet, Q. Liao, and Y. Zhu, "Kernel feature selection to fuse multi-spectral mri images for brain tumor segmentation," *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 256–269, 2011.

[15] A. Phinyomark, P. Phukpattaranont, and C. Limsakul, "Feature reduction and selection for emg signal classification," *Expert Systems with Applications*, vol. 39, no. 8, pp. 7420–7431, 2012.

[16] C.-P. Lee and Y. Leu, "A novel hybrid feature selection method for microarray data analysis," *Applied Soft Computing*, vol. 11, no. 1, pp. 208–213, 2011.

[17] H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," *Knowledge-Based Systems*, vol. 24, no. 7, pp. 1024–1032, 2011.

[18] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt, "Feature selection via dependence maximization," *The Journal of Machine Learning Research*, vol. 98888, no. 1, pp. 1393–1434, 2012.

[19] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, "Use of the zero norm with linear models and kernel methods," *The Journal of Machine Learning Research*, vol. 3, pp. 1439–1461, 2003.

[20] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Advances in neural information processing systems*, 2005, pp. 507–514.

[21] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," *The Journal of Machine Learning Research*, vol. 5, pp. 845–889, 2004.

[22] Z. Xu, I. King, M.-T. Lyu, and R. Jin, "Discriminative semi-supervised feature selection via manifold regularization," *Proceedings of IEEE Transactions on Neural Networks*, vol. 21, no. 7, pp. 1033–1047, 2010.

[23] Z. Zhao and H. Liu, "Semi-supervised feature selection via spectral analysis." in *SDM*. SIAM, 2007, pp. 641–646.

[24] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial intelligence*, vol. 97, no. 1, pp. 245–271, 1997.

[25] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1, pp. 273–324, 1997.

[26] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani *et al.*, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[27] J. R. Quinlan, *C4. 5: programs for machine learning*. Morgan kaufmann, 1993, vol. 1.

[28] U. M. Fayyad and K. B. Irani, "On the handling of continuous-valued attributes in decision tree generation," *Machine learning*, vol. 8, no. 1, pp. 87–102, 1992.

[29] Q. Zhu, L. Lin, M.-L. Shyu, and S.-C. Chen, "Feature selection using correlation and reliability based scoring metric for video semantic detection," in *Proceedings of 2010 IEEE Fourth International Conference on Semantic Computing (ICSC)*. IEEE, 2010, pp. 462–469.

[30] Q. Zhu, L. Lin, M.-L. Shyu, and S.-C. Chen, "Effective supervised discretization for classification based on correlation maximization," in *Proceedings of 2011 IEEE International Conference on Information Reuse and Integration (IRI)*. IEEE, 2011, pp. 390–395.

[31] L. Lin, M.-L. Shyu, and S.-C. Chen, "Association rule mining with a correlation–based interestingness measure for video semantic concept detection," *International Journal of Information and Decision Sciences*, vol. 4, no. 2, pp. 199–216, 2012.

[32] L. Lin and M.-L. Shyu, "Effective and efficient video high-level semantic retrieval using associations and correlations," *International Journal of Semantic Computing*, vol. 3, no. 04, pp. 421–444, 2009.

[33] L. Lin and M.-L. Shyu, "Weighted association rule mining for video semantic detection," *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, vol. 1, no. 1, pp. 37–54, 2010.

[34] 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 Tenth IEEE International Symposium on Multimedia, 2008. ISM'09*. IEEE, 2008, pp. 316–321.

[35] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen, "Effective feature space reduction with imbalanced data for semantic concept detection," in *2008*

IEEE on Sensor Networks, Ubiquitous and Trustworthy Computing, SUTC'08.International Conference on. IEEE, 2008, pp. 262–269.

[36] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen, "Video semantic concept discovery using multimodal-based association classification," in *2007 IEEE International Conference on Multimedia and Expo*. IEEE, 2007, pp. 859–862.

[37] L. Lin, C. Chen, M.-L. Shyu, and S.-C. Chen, "Weighted subspace filtering and ranking algorithms for video concept retrieval," *IEEE on MultiMedia*, vol. 18, no. 3, pp. 32–43, 2011.

[38] M.-L. Shyu, Z. Xie, M. Chen, and S.-C. Chen, "Video semantic event/concept detection using a subspace-based multimedia data mining framework," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 252–259, 2008.

[39] L. Lin, M.-L. Shyu, and S.-C. Chen, "Enhancing concept detection by pruning data with mca-based transaction weights," in *Proceedings of the 11th IEEE International Symposium on Multimedia, 2009. ISM'09.* IEEE, 2009, pp. 304–311.

[40] R. C. Prim, "Shortest connection networks and some generalizations," *Bell system technical journal*, vol. 36, no. 6, pp. 1389–1401, 1957.

[41] E. F. Moore, *The shortest path through a maze*. Bell Telephone System., 1959.

[42] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *Proceedings of the ACM international conference on image and video retrieval*. ACM, 2009, p. 48.

[43] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.