Exploiting Synergistic and Redundant Features for Multimedia Document Classification

Jana Kludas, Eric Bruno and Stephane Marchand-Maillet

University of Geneva, Switzerland kludas|bruno|marchand@cui.unige.ch

Summary. The task of multimedia document classification is challenging due to a diverse set of problems like a high dimensional, sparse and noisy features space, the unknown relevance of the features towards the classification target and the semantic gap between non-informative, low-level features and high-level semantic meanings. As a solution we propose a classification approach combined with feature selection and construction based on feature information interactions. This information-theoretic dependence measure can detect complex feature dependencies in multivariate settings. They help to find relevant and non-redundant features and hence allow efficient classification. Experiments on artificial and real world data show the superiority of feature selection based on N-way interactions over greedy, pair-wise dependence measures like correlation and mutual information.

Key words: multimedia document classification, feature selection, feature information interaction.

1 Introduction

Multimedia data processing received in the last decade a lot of attention by the research communities due to the 'multimediatisation' of the World Wide Web as well as other data collections in allday life. The most important problems identified in multimedia-based classification and retrieval are, amongst others, the high dimensionality, sparseness and noisiness of the multi modal feature space, the unknown relevance of features and modalities towards the classification target and the semantic gap between low-level features and high level semantic meanings.

In theory, classifiers that use more features have more information and thus should have more discriminating power. But in practice, the curse of dimensionality is degrading classification results, partly because the actual relevant information is obscured by many irrelevant features [12]. The influence of redundant features is even worse as was shown in [14], because they cause over-fitting of the data. So, one can apply a feature selection algorithm that

finds an optimal subset (where optimality is still under discussion) of maximal relevant and minimal redundant features. Feature selection also helps to reduce the computational complexity, memory usage and number of training examples needed in large scale applications like in multimedia document classification.

Another big problem that was already mentioned above is the usage of non-informative, generic features. In [16] the authors show that class-specific features are superior to generic ones like wavelet components and color histograms, because they carry more information about the target problem. As a consequence, less complex classification approaches are needed, but this comes along with a loss in generality of the system. In case of generic, low-level features, no feature is informative by itself, but a group of features can be. Then, similarity-based learning algorithms that use the full feature set will fail, because the proximity in the instance space is not related to classification in this domain.

We propose an approach that accounts for these problems: exploitation of feature information interactions for feature selection and construction towards a more efficient information fusion and hence improved multimedia document classification. This information-theoretic dependence measure finds the exact, irreducible attribute interactions in a multivariate feature subset. For subsets of size $N = 2$ the interactions result in the well known mutual information, for higher order subsets $N > 2$, feature information interaction results in positive values that indicate synergy and negative values that indicate redundancy. The synergistic feature subsets are highly relevant to the classification target and minimal redundant, thus good classification results can be achieved by using only a small part of the full feature set.

In the next Section 2 we review some related work in feature selection and more specifically works that apply interactions. Then the theory on feature information interaction is given in 3, followed by the experimental section 4, where tests on artificial and real world data show the advantages as well as the problems of feature interactions.

2 Related Work

Due to the increasing demand in dimensionality reduction caused by larger and larger data collections, research in feature selection has seen extensive efforts in the last years. For an overview of feature selection methods we refer to [14].

In the early years of feature selection pure relevance based feature selection was performed, which means that each single feature was evaluated towards its relevance to the class label. Soon researchers found the pair-wise evaluation that ignores the multivariate setting insufficient [18], because the result set contained redundancy that hurts the classification result. So the interest

moved on to multivariate evaluation measures and subset search, where the biggest problem is the exponential search space.

The most acknowledged definition of relevance of features in the multivariate setting is defined based on Markov Blankets [14]. In a Bayesian network a set of nodes M that shield a given node Y from the influence of another node X builds a Markov Blanket, that means the features M are strongly relevant, the features connected to Y but not being on M are weakly relevant and all others X are irrelevant. This definition has though only value in theory since the exhaustive search in real world applications is prohibitive. But many methods emerged in the last years that try to approximate the Markov Blanket calculation or the idea of an optimal subset by maximizing relevance and minimizing redundancy like in fast correlation-based FS (FCBF) [18], ReliefF [13], tree dependent component analysis (TCA) [1], Rosetta [3], conditional mutual information maximization (CMIM) [6] and MinRedMaxRel [5] to name just a few.

Only since recently, interacting features were considered for feature selection and dimensionality reduction to defy heuristic-based methods. For example, [7] gives a good introductory discussion about the relevance of interactions for attribute construction, detection of Simpson's paradox, coping with attribute disjoints and why greedy attribute selection does not work well. In [19] the authors develop INTERACT, a feature selection system that finds interacting features based on a consistency measure. Contrary to an evaluation based on mutual information, the inconsistency measure is monotonic and hence allows an efficient search. Feature selection based on joint mutual information is presented in [17], but here it is only used to eliminate redundancies in the feature subset.

Probably the work that is most related to ours is [8]. In this bio-informatics application, they already use feature interactions, here synergistic gene pairs, to improve micro array-based classification, which needs simultaneous profiling of thousands of genes with various conditions. Since the problem's definition is very similar to our multimedia document classification task (high dimensionality, small number of training data, low level features) we got inspired to investigate its application in multimedia settings.

3 Feature Information Interaction

Interaction is a multivariate, information-theoretic based feature dependence measure [10, 9]. Before its introduction there was no unifying definition of feature dependence in multivariate settings, but similar formulas have emerged independently in other fields from physics to psychology. Feature information interaction or co-information as it was named in [2] is based on McGill's multivariate generalization of Shannon's mutual information. It describes the information that is shared exclusively by all of k random variables, without over counting redundant information in the pairwise attribute subsets. Thus it

finds irreducible and unexpected patterns in data that are necessary to learn from data [15].

The k-way interaction information as found in [10] for a subset $\mathcal{S}_i \subseteq \mathcal{X}$ of all attributes $\mathcal{X} = \{X_1, X_2, ..., X_n\}$ is defined as:

$$I(\mathcal{S}) = -\sum_{\mathcal{T}\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{T}|} H(\mathcal{T}) = I(S \setminus X|X) - I(\mathcal{S} \setminus X), X \in \mathcal{S} \quad (1)$$

with the entropy being $H(\mathcal{X}) = -\sum_{\mathcal{X}\in\mathcal{S}} P(X) log_2 P(X)$. In case of several variables the joint probability distribution is used. The feature interaction for $k = 1$ reduces to the single entropy, for $k = 2$ to the well known mutual information and for $k = 3$ attributes A, B, C to McGill's multiple mutual information:

$$I(A; B) = H(A) + H(B) - H(A, B) \quad (2)$$

$$\begin{aligned} I(A; B; C) = & \quad I(A; B|C) - I(A; B) \\ = & \quad H(A, B) + H(A, C) + H(B, C) \\ & - H(A) - H(B) - H(C) - H(A, B, C). \end{aligned} \quad (3)$$

According to this definition 3-way information interaction will be only zero iff A and B are conditionally independent in the context of C, because then $I(A; B|C) = I(A; B)$. So it gives the information exclusively shared by the involved attributes. Hence, information interactions are stable and unambiguous, since adding new attributes is not changing already existing interactions, but only adding new ones. Furthermore they are symmetric and undirected between attribute sets.

An important characteristic of k-way feature information interactions with $k > 2$ is that it can result in positive and negative values. Normally, when we consider Markov chains $A \rightarrow B \rightarrow C$, the data processing inequality states that conditioning always reduces the information $I(A; B|C) <= I(A; B)$. This way the 3-way mutual information would be limited to $I(A; B; C) \leq 0$. But the problem of feature interactions is not a Markov chain, that is why it is possible that $I(A; B; C) > 0$. For example, let $C = A + B$ and let A and B be independent random variables, then $I(A; B) = 0$, but $I(A; B|C) = H(A|C) - H(A|B, C) = P(C = 1)H(A|C = 1) = 0.5bit$. The variables A, B are said to have a synergy towards C. Thus, we can distinguish two different types of feature information interactions:

Synergy $I(A; B; C) > 0$: In case of positive interactions the process benefits from an unexpected synergy in the data. Synergy occurs when A and B are statistical independent, but get dependent in the context of C. If C is the class label A, B are relevant and non-redundant features, and hence build

an optimal feature subset. Greedy feature selection algorithms are unable to detect synergies in the data.

Redundancy $I(A;B;C) < 0$**:** Negative interactions occur when attributes partly contribute redundant information in the context of another attribute, which leads to a reduction of the overall dependence. If C is the class label A, B are relevant to the class, but the feature subset suffers from redundancy.

3.1 Approximation of 3-way Feature Information Interaction

The calculation of the full feature information interaction matrix is not feasible even for a rather small collections, since the size of all possible combinations is M^k, where k is the size of the feature subset and M the number of features. Alternatively, heuristic search or sampling strategies are used to approximate the result as it was done in many methods in Section 2.

We formalize our problem as follows: $d = [1, ..., N]$ are the multimedia documents, $f_d = [1, .., M]$ are the extracted low-level features and $l_d = [1, ..., C]$ are the class labels, that we have given as ground truth. The features and labels are represented as probabilities over the instances such as:

$$P(f^i_{d_j}) = \frac{m_{f_i}}{m_{d_j}} \text{ with } \sum_i P(f^i_{d_j}) = 1 \ \forall d_j. \tag{4}$$

and

$$P(l^i_{d_j}) = \begin{cases} 1, & d_j \in c_i \\ 0, & otherwise \end{cases} \tag{5}$$

where m_{f_i} is the number of occurrences of a feature in a document and m_{d_j} the number of all features occurring in a document. These descriptions are conform to the frequentist interpretation of probability and result in the discrete probability matrix $P(F)$ of size $[M \times N]$ and $P(L)$ of size $[C \times N]$.

We applied in the following experiments a sub-sampling strategy to approximate interactions with $k = 3$ between two features and the class label. Using a normal random distribution, we chose to draw a very small sample set $S << [MxMxC]$ from the original search space. To do so, we approximated the joint entropies of the features and class labels by contingency tables. Then, the approximated feature interaction $I_S(A;B;C)$ between two random features and the class label is calculated as described in the previous subsection.

This is in any case a sub optimal solution, since a lot of significant interaction will be missing. Furthermore, also a lot of redundant interaction values are calculated, which is due to the symmetry of the interactions.

Interactions-based Feature Selection In a filter approach for feature selection a feature subset F' of size M' is chosen as classifier input. The evaluation measure based on the 3-way feature information interactions $I(F_{k=2}, L)$ divides into the absolute value $argmax|I(F_{k=2}, L)|$ labeled as (abs), the synergies $argmax(I(F_{k=2}, L))$ (syn) and the redundancies $argmin(I(F_{k=2}, L))$ (red). As dependency measures to compare our approach with, we chose absolute correlation $argmax|Cor(F, L^T)|$ (corr), mutual information $argmaxI(F, L)$ (mut2D) and random selection (rand).

4 Experiments

In this section we present classification experiments that compare our approach of feature selection and construction based on feature information interaction to a baseline system that uses no feature selection. Furthermore, we compare our results to systems that perform a feature selection based on correlation and mutual information. The goal is to determine, if the knowledge of synergies and redundancies can improve a classification task and how it is best exploited.

4.1 Artificial Data Sets

First, we conducted feature selection tests on simple artificial data sets, where the functional relations between the input variables as well as their relations towards the class labels are known. The artificial, binary feature sets of size $N = 10.000$ were created as follows:

- AND problem: $l = 1$ if $(f^1 \bigwedge f^2 \bigwedge f^3)$, features f^4, f^5, f^6 are randomly distributed
- Parity problem: $l = 1$ if $oddNumber(f^1, f^2, f^3)$, features $f^{4..12}$ are randomly distributed
- ParityAND problem: $l = 1$ if $oddNumber(f^5, f^6) \bigwedge oddNumber(f^7, f^8)$, features $f^{1..4}, f^{9..12}$ are randomly distributed

The feature selection is performed as given in the previous section. For the artificial data sets the interactions of size $k > 2$ were computed exhaustively, since the small problem size allows this.

Table 1. Feature Selection based on random, correlation, mutual information and interactions

	corr	mut2D	mut3D(syn/red)	mut4D(syn/red)
AND	**2, 3, 1**	**2, 3, 1**	1, 3/**2, 3, 1**	**3, 1, 2**/−
Parity	−	−	1/−	**1, 3, 2**/2
ParityAND	5, 6	5, 6	**7, 8, 5, 6**/−	6, 7, 8/7

Table 1 shows the features that were selected by the pairwise and interaction-based evaluation measures. It is clearly visible that the greedy approaches, no matter if they are based on a dependence (correlation) or information measure (mutual information), can detect only a functional dependency that is based on AND connections and hence defines redundancy. The AND problem is as well solved by the redundant 3-way and the synergistic 4-way interactions.

In contrary, the interaction-based evaluation measures can detect features that have a complex OR relationship with the class label. For the Parity problem only the synergistic 4-way interactions find all features that are relevant to the problem. This is explainable by the fact that it is a $k = 4$ functional dependence of three variables and the class label. For the final ParityAND problem the situation is different. Here, already the 3-way interactions find all relevant features, whereas the 4-way interactions fail. The ParityAND problem is in fact constructed by two $k = 3$ problems, such that they can best be detected by the 3-way interaction.

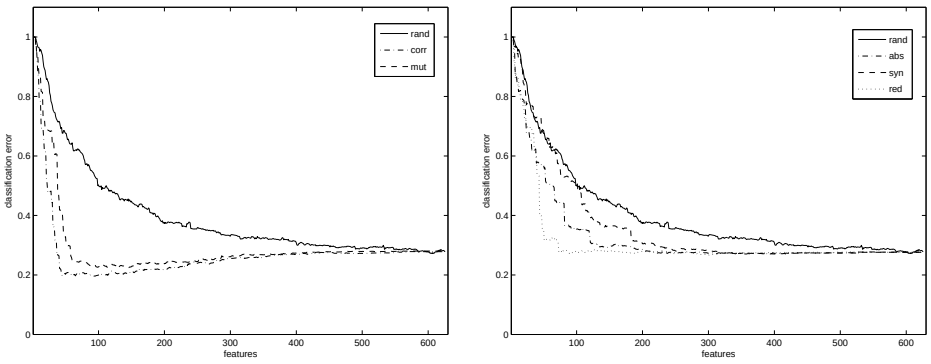
4.2 Real Data Set

For the real data experiments, we used the Washington collection, which consists of $N = 886$ documents, which are images annotated with 1 to 10 keywords. They are grouped into $C = 20$ classes like for example football, Barcelona and swiss mountains. The extracted feature set F consists of the global color and texture histograms which result in $M_c = 165$ and $M_x = 164$ features respectively. Additionally, we constructed a textual feature vector of size $M_t = 297$ with the term frequencies of the keywords. The continuous variables are discretized with a simple equal length quantizer.

The classification is done with the SVM light library [11] using a RBF kernel. We followed two different strategies: (1) a late or hierarchical fusion over each modality and (2) late fusion over feature subsets of size $k = 3$. We ran a cross validation to optimize the parameters of the SVM. As training set we randomly selected for each run and class 5 positive and 7 negative examples, the rest of the examples were used as test set. The experiments were run with the one-against-all classification strategy, where the classification errors were averaged over all classes and over 10 runs.

Fusion at modality level For this strategy we sort the selected features into their modalities. Then, one SVM is applied to every modality (color, texture, text) and their results are again combined with a SVM. Figure 1 shows the classification errors of the random, correlation and mutual information based feature selection at the left hand side and the for the interactions on the right hand side.

Correlation performs best with $e = 0.19$ at $M' = 46$, mutual information is worse in terms of the classification error as well as the feature set size with $e = 0.22$ with $M' = 100$ features. But both outperform the classification error that is achieved with the full set $e = 0.28$. Compared to this the sub sampled, redundant 3-way interactions perform only slightly better than the full set



(a) random, correlation, mutual information (b) 3-way absolute, synergistic and redundant interaction

Fig. 1. classification errors for the Washington collection for fusion at modality level

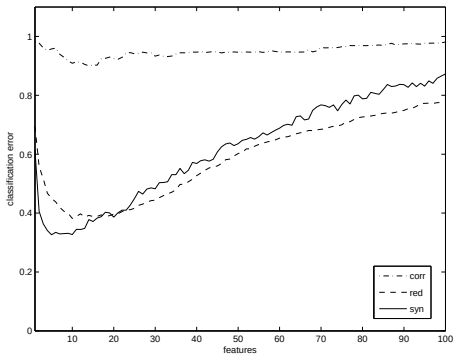
with $e = 0.27$ at $M' = 77$ as seen in Figure 1(b). We think that this is due to the inefficient sub sampling strategy that misses many of the important feature relations. Synergistic interactions are outperformed by the full set, which, we think, showd that synergies can not be exploited with standard feature selection approaches.

Fusion at feature subset level It is for this reason that we tried a simple feature construction approach on the same data and feature evaluation measures. It is set up again as a hierarchical SVM. But now, we create in the first step a mid-level feature over each synergistic, redundant or correlated feature subset by using a SVM. The results are then fused in a second level SVM towards the final classification result.

The classification errors are shown in Figure 2. Now the synergistic features outperform largely the redundant ones with $e = 0.32$ at only $M' = 5$ feature subsets of size $k = 3$, hence it uses only 10 features plus the class label. Still, the synergistic feature subsets stay behind the performance of the full feature set, but it achieves an acceptable classification result with only $1/100$ of the original feature set. This is the steepest reduction of the classification error within the first few features, which makes this strategy valuable for extreme feature selection.

5 Conclusions and Future Work

We presented the feature information interaction as an evaluation measure in feature and feature subset selection to improve large scale multimedia document classification. Since the full calculation of the interaction matrix can not be calculated we propose a sub sampling strategy for its approximation.



(a) 3-way synergistic and redundant interaction, correlation

Fig. 2. classification errors for the Washington collection for fusion at feature subset level

From the artificial data experiments, we can conclude that feature interactions can detect complex functional dependencies between features and class labels, which can help to improve classification results. Pairwise, greedy evaluation measure fail, if the functional feature dependence comprises an OR connection or higher feature hierarchies. These results were achieved with a complete calculation of the interaction matrix, which was feasible for the small problem size.

In the real data experiments, where the full calculation was impossible, the feature information interaction can not outperform the standard pairwise correlation-based feature selection. We hope to overcome this performance descent by developing a more efficient calculation of the feature interactions, that finds all relevant interactions. Then, the fusion at subset level of the synergistic features can be a promising approach for extreme feature selection with only a little loss in performance.

6 Acknowledgments

This work is funded by the European project MultiMATCH (EU-IST-STREP#033104).

References

1. Kobus Barnard, Pinar Duygulu, and David Forsyth. Recognition as translating images into text. In *Internet Imaging IX, Electronic Imaging*, 2003.
2. A.J. Bell. The co-information lattice. In *4th Int. Symposium on Independent Component Analysis and blind Signal Seperation (ICA2003)*, pages 921–926, 2003.

3. B. Blum, M. I. Jordan, D. Kim, R. Das, P. Bradley, R. Das, and D. Baker. Feature selection methods for improving protein structure prediction with rosetta. In *Advances in Neural Information Processing (NIPS)*. MIT Press, 2008.
4. Manoranjan Dash, Huan Liu, and Hiroshi Motoda. Consistency based feature selection. In *PADKK '00: Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*, pages 98–109. Springer-Verlag, 2000.
5. Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. In *CSB '03: Proceedings of the IEEE Computer Society Conference on Bioinformatics*, page 523. IEEE Computer Society, 2003.
6. Francois Fleuret. Fast binary feature selection with conditional mutual information. *Machine Learning Research*, 5:1531–1555, 2004.
7. Alex. A. Freitas. Understanding the crucial role of attribute interaction in data mining. *Artificial Intelligence Review*, 16(3):177–199, 2001.
8. B. Hanczar, J.-D. Zucker, C. Henegar, and L. Saitta. Feature construction from synergic pairs to improve micro-array classification. 23(21):2866–2872, 2007.
9. A. Jakulin and I. Bratko. Analyzing attribute dependencies. In *Principles of Knowledge Discovery in Data (PKDD)*, pages 229–240, 2003.
10. A. Jakulin and I. Bratko. Quantifying and visualizing attribute interactions, 2003.
11. Thorsten Joachims. *Learning to Classify Text Using Support Vector Machines, Dissertation*. Kluwer, 2002.
12. Daphne Koller and Mehran Sahami. Toward optimal feature selection. In *International Conference on Machine Learning*, pages 284–292, 1996.
13. I. Kononenko, E. Simec, and M. Robnik-Sikonja. Overcoming the myopia of inductive learning algorithms with relieff. *Applied Intelligence*, 7(1):39–55, 1997.
14. Huan Liu and Hiroshi Motoda. *Computational methods of Feature Selection*. Chapman & Hall/CRC, 2008.
15. I. Perez. *Learning in presence of complex attribute interactions: An Approach Based on Relational Operators*. PhD thesis, 1997.
16. Michel Vidal-Naquet and Shimon Ullman. Object recognition with informative features and linear classification. In *Ninth IEEE International Conference on Computer Vision (ICCV'03)*, pages 281–290, 2003.
17. Hua Howard Yang and John Moody. Feature selection based on joint mutual information. In *Advances in Intelligent Data Analysis (AIDA)*, 1999.
18. Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *The Twentieth International Conference on Machine Learning (ICML-2003)*, 2003.
19. Zheng Zhao and Huan Liu. Searching for interacting features. In *International Joint Conference on Artificial Intelligence*, page 1156, 2003.