Up-Fusion: An Evolving Multimedia Decision Fusion Method

Xiangyu Wang
National Univ. of Singapore
wang06@comp.nus.edu.sg

Yong Rui
Microsoft China R&D Group
yongrui@microsoft.com

Mohan S. Kankanhalli
National Univ. of Singapore
mohan@comp.nus.edu.sg

ABSTRACT

The amount of multimedia data available on the Internet has increased exponentially in the past few decades and is likely to keep on increasing. Given multimedia's nature of having multiple information sources, fusion methods are critical for its data analysis and understanding. However, most of the traditional fusion methods are static with respect to time. To address this, in recent years, several evolving fusion methods have been proposed. However, they can only be used in limited scenarios. For example, the context aware fusion methods need the context information to update the fusion model, but the context information may not always be available in many applications. In this paper, a new evolving fusion method is proposed based on the online portfolio selection theory. The proposed method takes the correlation among different information sources into account, and evolves the fusion model when new multimedia data is added. It can deal with either crisp or soft decisions without requiring additional context information. Extensive experiments on concept detection tasks over TRECVID dataset have been conducted, and very promising results have been obtained.

Categories and Subject Descriptors

H.3.1 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis and Indexing; H.1.0 [**MODELS AND PRINCIPLES**]: General

1. INTRODUCTION

Multimedia techniques are developing at an unprecedented pace. The amount of multimedia data available on the Internet has increased exponentially. Analysis of multimedia data is therefore needed in many applications such as information retrieval, education, and security. To perform multimedia analysis tasks, fusion methods are often employed.

There are still some open issues in multimedia fusion. One important issue is that the fusion model is not evolving. The evolution of the fusion model is of primary importance because of the nature of multimedia applications. First of all, most of multimedia data has limited or no labeled information. For example, on Flickr, the label for the multimedia document (image, tags and description) is not available or quite noisy. The semantic label is important for multimedia analysis because many multimedia analysis tasks are based on classification and a large amount of labeled training data is necessary for good classification. Labeled examples are fairly expensive to obtain due to the high labor costs faced when annotating videos. Thus, little amount of training data is available at the beginning. The fusion performance may suffer as a result. Furthermore, the multimedia data keeps increasing with time. New instances of multimedia data is continuously added. For example, new videos are periodically uploaded on Youtube. The nature of the data collection can change. Thus, the fusion model may not always be valid or effective as the multimedia data increases. It will be quite useful to evolve the fusion model and improve the performance with new data. The previous methods generally cannot cope with the new data well. In this paper, an evolving fusion method, called Up-Fusion, is proposed.

MM'11, November 28–December 1, 2011, Scotsdale, Arizona, USA.

2. RELATED WORKS

Most of the traditional decision fusion methods are static fusion methods [1]. That is, the fusion models in the methods stay unchanged no matter how the nature of data varies. For example, max / min / average fusion takes the max / min / average decision score of all information sources as the final decision score. The training-based super-kernel fusion method is proposed by Wu *et al.* in [9]. Not merely training on individual information source to acquire individual classification model, the method determines the optimal combination of information sources by further training on the output decision scores of different information sources.

Generally speaking, the correlation and different performances of information sources are generally not considered. By considering correlation, a fusion method based on the portfolio selection theory is proposed in [8]. With the mean-variance analysis, the portfolio fusion finds the optimal weights for different sources by minimizing the correlation while maximizing the performance. But it is still a static method.

More importantly, once obtained, the fusion models in these fusion methods are static over time. In reality, the correlation and reliability of information sources might vary with the changes of data or context. The static fusion methods cannot adapt to the changing data and environment, which may make the methods unreliable or even fail to work.

Particularly, the portfolio fusion method [8] cannot be simply extended for evolution: simply applying portfolio fusion cannot guarantee to improve the fusion performance and it is inefficient to update fusion model whenever there is a new data instance.

Several evolving fusion methods have been proposed. An adaptive *crisp decision fusion method* is proposed in [4]. They modeled the decisions as conditional probabilities and used log-likelihood as weights for individual source. The weights are updated according to their agreements with the fusion decision at each iteration. However, only crisp decisions (e.g., "yes / no") are considered, and it is not consistent with Principle of Least Commitment. The possible hypotheses are dropped intermediately and the performance may be degraded. A *confidence evolution method* is described in [2]. The method needs training for initial confidence for individual source. Then, at each instance, the sources are divided into two subsets based on their decisions. The confidences are updated according to their agreement coefficients with the subsets. The methods need trusted sources and only confidence is updated. The fusion model is based on the underlying assumption that the media streams are independent, and the correlation among sources are not considered. The method needs to update the confidence for each new instance. It will be inefficient, and a significant restriction is that the labels may not be available online, as it may require manual intervention at every update step. A more realistic scenario is the update of the existing fusion model when a new batch of data becomes available. Recently, some context aware fusion methods have been proposed like [7, 5]. In *context weight fusion method* [7], adaptive weighting scheme was adopted for acoustic and visual speech recognition. The weights for audio and visual vary according to the noise level in speech. The method needs the context information which may not be available and dealing with all influential context factors is unrealistic. Again, correlation among information sources are not considered.

In this paper, we propose an evolving fusion method based on the online portfolio selection theory. Online portfolio selection [6] is a mechanism developed in economics. Consider a portfolio containing n stocks. Each trading day, the performance of the stocks can be described by a vector of price relatives, denoted by $\mathbf{x} = \{x_1, \ldots, x_n\}$, where x_i is the next day's opening price of the ith stock divided by its opening price on the current day. A portfolio is defined by a weight vector $\mathbf{w} = \{w_1, \ldots, w_n\}$ such that $w_i \geq 0$ and $\sum_{i=1}^{n} w_i = 1$. w_i is the proportion of the total portfolio value invested in the ith stock. The online portfolio selection strategy is as follows: At the start of each day t, the strategy gets the previous price relatives $\mathbf{x}^1, \ldots, \mathbf{x}^{t-1}$. From this information, the strategy immediately selects its portfolio $\mathbf{w}^t$ for the day. Over time, a sequence of daily price relatives $\mathbf{x}^1, \ldots, \mathbf{x}^T$ is observed and a sequence of portfolios $\mathbf{w}^1, \ldots, \mathbf{w}^T$ is selected. The mechanism aims to maximize the wealth on each day based on previous observations. Similarly, we want to improve the multimedia fusion performance as the data increasing in multimedia systems. Everyday we can observe the price and the return in the stock investment. In multimedia fusion, the scenario is similar if the "correct" labels of the new instances can be revealed for each update.

Compared to the static fusion methods, our proposed method is evolutionary. The fusion model evolves as new data being added. In this way, the fusion model can adapt to the changing data and environment. Suitable fusion models for different conditions should improve the performance than a fixed model. Compared to the previous evolving fusion methods, our proposed method utilizes the correlation among different information sources, can deal with either crisp or soft decision (confidence score), and no context information is required.

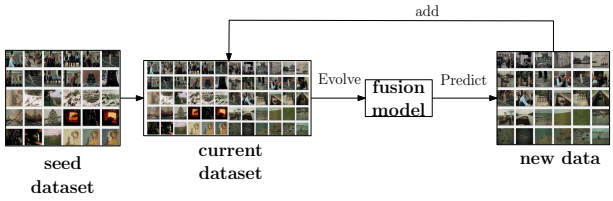

Figure 1: Framework of proposed Up-Fusion method

3. UP-FUSION METHOD

$\mathcal{S}$ is a multimedia system designed for performing a task D. It consists of $n \geq 1$ correlated information sources $M_1, \ldots, M_n$. For $1 \leq i \leq n$, let $I_i(\mathbf{X}) \in [0, 1]$ be the decision of the task D based on M_i on instance $\mathbf{X}$. It is usually obtained by employing a detector on the features extracted from M_i. The final prediction I of $\mathcal{S}$ is modeled as the fusion of $I_i(\mathbf{X}), i = 1, \ldots, n$ based on the fusion model. Let $r_i(\mathbf{X})$ be the return of M_i at $\mathbf{X}$, and R_i be the expected return of M_i, which is defined as $R_i = E[r_i]$. More specifically, $r_{i;X_{\alpha:\beta}}$ denotes the returns for instances $\mathbf{X}_\alpha$ to $\mathbf{X}_\beta$ based on M_i. For $1 \leq i, j \leq n$, let $\Phi = [\Phi_{ij}]$ be the covariance matrix of information sources. The element Φ_{ij} is defined as $\Phi_{ij} = E[(r_i - E[r_i])(r_j - E[r_j])]$. It captures the correlations of different information sources. For $0 \leq t \leq T$, let f_i^t be the classification model of M_i at iteration t, and F^t be the multimedia fusion model obtained at iteration t. $y(\mathbf{X})$ is the true label of instance $\mathbf{X}$. The label is the class the data instance belongs to. The fusion flow is described in Algorithm 1, and the procedure can be illustrated in Figure 1.

The definition of return can be varied to different applications according to their aims. In the classification problem, since the aim of the classifier of information source is to accurately predict the labels and the performance is evaluated using accuracy, the return should be positive if the prediction is correct and negative otherwise. For M_i on instance $\mathbf{X}$, the return is defined as:

$$r_i(\mathbf{X}) = \begin{cases} 1 & \text{if } h_i(\mathbf{X}) == y(\mathbf{X}) \\ -1 & \text{otherwise} \end{cases} \quad (1)$$

where $h_i(\mathbf{X})$ is the predicted class of M_i on instance $\mathbf{X}$. For the retrieval problem, where we evaluate the performance using average precision, the definition of return can be:

$$r_i(\mathbf{X}) = \begin{cases} I_i(\mathbf{X}) - 0.5 & \text{if } y(\mathbf{X}) == 1 \\ -(I_i(\mathbf{X}) - 0.5) & \text{otherwise} \end{cases} \quad (2)$$

The expected return of ith information source R_i is approximated as $R_i = E[r_i]$ over all the instances. The risk of information source is modeled as the standard deviation σ of return. For M_i, $\sigma_i^2 = E[(r_i - E[r_i])^2]$. The correlation ρ_{ij} between M_i and M_j over instances $\mathbf{X}_{\mathbf{1}:\mathbf{N}}$ is defined

Algorithm 1 Proposed Up-Fusion Method

Input: Seed dataset (the initial labeled dataset)
Initialization (Section 3.1)

- With the seed dataset, the classification model f_i for individual information source can be obtained
- The return $\mathbf{R}^0$, as well as the covariance matrix Φ^0 can be obtained according to Equation (3) and (4) based on the seed dataset
- The initial fusion model F^0 is constructed using Equation (5)

Evolution (Section 3.2)

- At iteration t, K new instances are added. The decisions can be obtained using the previous model F^{t-1}
- Consequently, the expectation $\mathbf{R}^t$ and correlation Φ^t for the information sources will be updated using Equation (6) and (7). The fusion model F^t will thus be updated according to Equation (8)

Output: Fusion model F^t

as: $\rho_{ij} = \frac{E[(r_i - E[r_i])(r_j - E[r_j])]}{\sigma_i \sigma_j}$. Thus, the covariance matrix for n information sources is $\Phi = [\Phi_{ij}]_{n \times n}$, in which $\Phi_{ij} = \rho_{ij}\sigma_i\sigma_j = E[(r_i - E[r_i])(r_j - E[r_j])]$

With the portfolio fusion method, the optimal weights $\mathbf{w}$ are obtained by minimizing $\mathcal{F} = \mathbf{w}^{\mathrm{T}}\Phi\mathbf{w} - \lambda\mathbf{R}^{\mathrm{T}}\mathbf{w}$. Here, $\mathbf{w}^{\mathrm{T}}\Phi\mathbf{w}$ is the risk of the information sources. $\mathbf{R}^{\mathrm{T}}\mathbf{w}$ is the expected return. $\lambda \in [0, +\infty)$ is a "risk tolerance" factor.

3.1 Initialization

The method starts with a dataset of N_0 labeled instances. This dataset is called the *seed dataset*. The classification model for individual information source can be obtained with the labeled data. Here, binary classification is considered because multi-class classification can be achieved by One-Versus-the-Rest strategy. The classification model for M_i is denoted as f_i^0. The decision according to f_i^0 on instance $\mathbf{X}$ is $I_i(\mathbf{X})$.

With the initial dataset, the expected return $\mathbf{R}^0$ and covariance Φ^0 are calculated. The initial expected return is

$$\mathbf{R}^0 = [R_i^0]_{n \times 1} \tag{3}$$

The initial covariance matrix for n information sources is $\Phi^0 = [\Phi_{ij}^0]_{n \times n}$, in which

$$\Phi_{ij}^0 = \rho_{ij}^0 \sigma_i^0 \sigma_j^0 \tag{4}$$

The optimal weights $\mathbf{w}^0$ for each information source are obtained by minimizing

$$\mathcal{F} = (\mathbf{w}^0)^{\mathrm{T}}\Phi^0(\mathbf{w}^0) - \lambda(\mathbf{R}^0)^{\mathrm{T}}(\mathbf{w}^0) \tag{5}$$

The initial fusion model $F^0 = \mathbf{w}^0 \cdot f_i^0$.

3.2 Evolution

The fusion model is updated every iteration when new data is added. It will be inefficient to update the fusion model whenever there is a new data instance. Moreover, a significant constraint is that the labels will not be discovered soon after the prediction is made. In our Up-Fusion method, we will update the fusion model when a batch of K new instances becomes available. In iteration $t (t = 1, 2, \ldots, T)$, K new instances are added into the dataset and the data instances are $\mathbf{X}_{1:N_t}$.

According to the definition, the return $\mathbf{R}^t = [R_i^t]_{n \times 1}$, in which R_i^t is defined as:

$$R_i^t = E[r_{i;X_{\alpha_t:\beta_t}}] \tag{6}$$

The covariance Φ_{ij}^t between M_i and M_j is updated as

$$\Phi_{ij}^t = \rho_{ij}^t \sigma_i^t \sigma_j^t = E[(r_{i;X_{\alpha_t:\beta_t}} - R_i^t)(r_{j;X_{\alpha_t:\beta_t}} - R_j^t)] \tag{7}$$

Here, the exact return and covariance method is used. That is, take all the current available data instances $\mathbf{X}_{1:N_t}$ into account, and calculate the return on the instance with Equation (1) or (2). Then, the new R_i^t and Φ^t is re-calculated on the whole available dataset based on the Equation (6) and (7). Here, $\alpha_t = 1$ and $\beta_t = N_t$.

The distribution of the newly added data instances may be largely different from the actual distribution, or the correlation of information sources on the newly added data instances varies from the actual correlation. The noisy new data instances may degrade the fusion performance. Thus, merely computing the exact return and covariance may not always improve the results. The performance may be unstable as the data increasing. To overcome this disadvantage, we refine the evolving fusion method by introducing a validation step. When the new data instances are added, the weights can be obtained with the Up-Fusion method. Then, the weights are validated on the initial seed dataset. If the performance on the initial seed dataset is improved compared to the previous weights, the new weights are updated. Otherwise, the weights remain unchanged. In this way, we can expect the fusion performance to be always improved.

Thus, the weights $\mathbf{w}^t$ at iteration t are obtained by minimizing

$$\mathcal{F} = (\mathbf{w}^t)^{\mathrm{T}}\Phi^t(\mathbf{w}^t) - \lambda(\mathbf{R}^t)^{\mathrm{T}}(\mathbf{w}^t) \tag{8}$$

Subject to:

- $\sum_{i=1}^{n} w_i^t = 1$, and $0 \leq w_i^t \leq 1$
- $\mathcal{P}(\mathbf{w}^t) \geq \mathcal{P}(\mathbf{w}^{t-1})$. Here, $\mathcal{P}(\mathbf{w})$ denotes the fusion performance on seed dataset with weights $\mathbf{w}$

To take the prior knowledge into account, the initial point for minimization is set to be the previous weights. Starting from the initial weight vector, the formula is optimized as a quadratic programming problem. If the performance on validation dataset with new weights is better than that of the old weights, the fusion model is updated with new weights. Otherwise, the weights keep unchanged. In this way, the method evolves the fusion model to improve the fusion performance. The fusion model at iteration t is then expressed as: $F^t = \mathbf{w}^t \cdot f_i^t$. Here, $f_i^t = f_i^0$ because the classification model is not re-trained when adding new data instances.

The evolution is one of our contributions. Compared to static fusion method, the evolution updates fusion model every iteration when new data is added. Compared to the previous evolving fusion methods, the evolution utilizes the correlation among different sources, can deal with either crisp or soft decision, and no context information is required.

4. EXPERIMENTS AND DISCUSSION

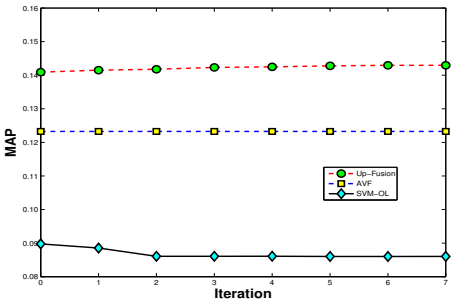

Figure 2: MAP based on whole exact return and covariance with true labels. Circle denotes the results of Up-Fusion method, square denotes average fusion method, while diamond denotes SKF-OL method

Methods	AVF	SKF	PTF	SKF-OL	Up-Fusion
MAP	0.123	0.09	0.141	0.086	0.143

Table 1: Performance of different fusion methods

To show the effectiveness of the proposed Up-Fusion method, experiments have been conducted on concept detection on TRECVID 2007 dataset. The concept detection is an important task in information retrieval. The performance is compared with the popular state-of-the-art fusion methods: average fusion method (AVF), super-kernel fusion method (SKF) and portfolio fusion method (PTF). For complete comparison, we give an online version of super-kernel fusion method by re-training the fusion model with SVM at each iteration, which is denoted as SKF-OL. For SVM training, LIBSVM [3] is used with RBF kernel and default parameter values. $\lambda = 1$ is used.

For the concept detection, the models are trained using three features: edge direction histogram, Gabor, and grid color moment [10]. There are 21,532 instances in the dataset. The data is evenly divided into three parts: initial part, new data part, and evaluation part. The initial part is taken as the initial seed dataset. The new data part is used to simulate adding new data instances. Then, we evaluate the performance for different concepts on the evaluation part of the dataset. In the evolution step, at each iteration, we sequentially include $K = 1,000$ instances from new data part into the available dataset and update the fusion models. Total 32 concepts are evaluated. The mean average precision (MAP) for all concepts is used as the performance criteria. Here, the average precision for each concept is calculated over the $2,000$ retrieved relevant shots.

The MAP for each iteration is shown in Figure 2. The MAP results of different fusion methods are given in Table 1. Compared to the MAP of average fusion method, which is 0.123, the final MAP for Up-Fusion method on whole data is 0.143. Compared to the portfolio fusion method that utilizes the initial dataset only and stays unchanged as data increases, the proposed Up-Fusion improve the performance by evolving the fusion models as new data is added. The Up-Fusion method improves PTF by 1.4%(relative). Compared to other fusion methods, the improvement is more obvious.

Generally speaking, the proposed method obtains better performance than the average fusion method and super-kernel fusion method. The evolution phase generally improves the results. However, the improvement is not quite much. It should be because the distribution and nature of the data in this experiment does not change much, so does the correlation between different information sources. Thus, the update of correlation in each iteration only slightly improve the performance because of more data. Surprisingly, the performance of the online super-kernel fusion method generally decreases when it takes the new data into account. It may be because the generalization performance tends to suffer when there is too much noise and unbalanced data.

5. CONCLUSIONS

In this paper, an evolving fusion method has been proposed. Compared to the previous static fusion methods, especially the portfolio fusion method, as new data is continually added, the proposed Up-Fusion method evolves to adapt to the changing data and environment conditions. Evolved fusion models for different conditions can perform better than a fixed fusion model. Compared to the previous evolving fusion methods, our method utilizes the correlation among different information sources, can deal with either crisp or soft decision, and no context information is required. Experiments on representative concept detection tasks have shown the superiority of the proposed Up-Fusion method. Better updating methods will be studied in the future.

6. REFERENCES

[1] P. K. Atrey, M. A. Hossain, A. E. Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, 16(6):345–379, 2010.

[2] P. K. Atrey and A. E. Saddik. Confidence evolution in multimedia systems. *IEEE Transanctions on Multimedia*, 10(7):1288–1298, November 2008.

[3] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[4] J.-G. Chen and N. Ansari. Adaptive fusion of correlated local decisions. *IEEE Transactions on Systems, Man, and Cybernetics*, 28(2):276–281, 1998.

[5] X. Geng, K. Smith-Miles, L. Wang, M. Li, and Q. Wu. Context-aware fusion: A case study on fusion of gait and face for human identification in video. *Pattern Recognition*, 43(10):3660–3673, 2010.

[6] D. P. Helmbold, R. E. Schapire, Y. Singer, and M. K. Warmuth. On-line portfolio selection using multiplicative updates. *Mathematical Finance*, 8(4):325–347, 1998.

[7] J.-S. Lee and C. H. Park. Adaptive decision fusion for audio-visual speech recognition. *Speech Recognition, Technologies and Applications*, pages 275–296, 2008.

[8] X. Wang and M. S. Kankanhalli. Portfolio theory of multimedia fusion. In *ACM International Conference on Multimedia*, pages 723–726, 2010.

[9] Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith. Optimal multimodal fusion for multimedia data analysis. In *ACM International Conference on Multimedia*, pages 572–579, 2004.

[10] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu. Columbia university's baseline detectors for 374 lscom semantic visual concepts. TechReport 222-2006-8, Columbia University, 2007.