Similarity Kernel Learning

Bálint Daróczy[1] Krisztian Buza[2] András A. Benczúr[1]
[1]Institute for Computer Science and Control, Hungarian Academy of Sciences (MTA SZTAKI)
[2]BioIntelligence Lab, Institute of Genomic Medicine and Rare Disorders, Semmelweis University
{daroczyb,benczur}@ilab.sztaki.hu, buza@biointelligence.hu

ABSTRACT

Kernel methods are popular in machine learning tasks. For Support Vector Machine classification or Support Vector Regression, the central question is the selection of the appropriate kernel. The task is difficult in particular if the data points have complex or multimodal attributes such as time series or visual content enhanced with geographic, numeric or text metadata. Unlike earlier approaches of the so-called Multiple Kernel Learning problem, where a large number of kernels are fused by wrapper methods as part of the optimization process, in this paper we mathematically derive an optimal kernel for the data set in question. We begin with selecting appropriate distances for the appropriate modalities, for example dynamic time warping distance for time series and Jensen-Shannon distance for the bag of words text representation. Our kernel is defined, without needs of wrapper methods, by considering the distances as attributes generated by a Markov Random Field. For the Markov Random Field, the natural kernel is based on the Fisher information matrix and its exact form can be computed from the data. We experiment with the above similarity kernel over a wide variety of data sets, including

- 64-channel EEG data;
- General time series data sets;
- Images with text annotations;
- Web documents;
- Gene expression levels.

Over the complex, multimodal or multiple time series classification tasks, our method outperforms the state of the art while reaching identical performance even over the simple unimodal problems as well, hence our method seems applicable under very general settings.

General Terms

Kernel methods, Classification, Mining rich data types, Similarity-based methods, Bioinformatics, Web mining

Keywords

Fisher information matrix, Support Vector Machine

1. INTRODUCTION

Kernel methods [47] are popular in various fields of data mining and knowledge discovery such as classification, regression, clustering or dimensionality reduction. Its numerous applications range from relation extraction [57] to the prediction of protein-protein interactions [4] and other problems in computational biology [45].

While kernel methods are well-founded from the theoretical point of view, the selection of the appropriate kernel is essential in many real-world tasks. In order to allow wide range of applications, various kernels have been introduced in the last decades such as the general-purpose polynomial and RBF-kernels as well as application-specific kernels, see e.g. string-kernels in text mining [6] or computational biology [31]. Learning optimal hyperparameters of these kernels may be computationally prohibitive in case of large datasets. Furthermore, even if the best hyperparameters have been found, the resulting kernel may not completely reflect the true structure of the data, which is likely to manifest in suboptimal results regardless of the particular analysis task.

The selection of feature set dependent distance or similarity metrics is crucial for learning. Although selecting and in some cases computing the potential metrics may constitute a challenging task, once metrics are defined, they can often be used to transform the original complex optimization problem to a less challenging one. Most notably, the Support Vector Machine (SVM) optimization phase is independent of the underlying metric based on precomputed kernel values.

An additional and interesting opportunity arise from the freedom of selecting similarity or distance metrics to define SVM kernels. In a number of practical applications such as image or document classification, we have to learn over multiple representations, often with different kernel functions. Images are often enriched by text description or other non-visual metadata such as geo-location or date, yielding a multimodal classification task with each mode (visual, text, geospatial) having its own natural metric [19]. Another example is Web content, where text, hyperlinks and style give us different kernels when categorizing Web pages or filtering Web spam [10].

In order to address the kernel selection problem, in this paper, we propose a principled meta-kernel learning approach based on Fisher information theory. Our new approach is computationally inexpensive and needs no wrapper methods

for learning a kernel over multiple modalities. In experiments on publicly-available real-world datasets from various domains such as classification of images, texts, time series and gene expression data, we show that our approach outperforms the state-of-the-art.

2. RELATED WORK

In many cases, one single kernel may perform suboptimally. In the last decade, this issue has primarily been addressed in the framework of multiple kernel learning (MKL) [3, 30, 49, 23]. With proposing a method to learn a kernel over multiple modalities, in this paper, we address a problem that is related to MKL, but is substantially different from MKL in several respects. First, we assume that all the modalities are used in the kernel, not only a fraction of them. Second, in order to devise a computationally efficient approach, we only calculate the distance between each instance and a small set of reference instances. This is in contrast to MKL techniques that require full kernel matrices. Last, but not least, our approach runs only one SVM optimization procedure while most MKL approaches are wrapper approaches and therefore they execute large amount of SVM optimizations.

Selecting the appropriate kernel under multiple modalities can be seen as a special case of the Multiple Kernel Learning problems where the kernels are computed on different feature sets. Bach et al. [39] suggested to solve the MKL problem with an iterative, wrapper like, sparse algorithm where in each iteration they solve a standard SVM dual problem and update the weights of the basic kernels. Instead of optimizing multiple times over the training set with a combination of kernel functions, we will define a novel kernel function combining all the representations into a single feature space. Our method is wrapper-free and is hence scalable for large data sets as well.

Late fusion approaches, see e.g. [56] and the references therein, combine the outputs of various kernel methods. Usually, they take an estimated certainty of each kernel method into account. In contrast to late fusion, our approach learns a kernel over various modalities instead of combining the outputs of different kernel methods.

3. THE SIMILARITY KERNEL

A natural idea to handle distances of pairs of observation is to use kernel methods. A kernel acts as an inner product between two observations in certain large dimensional space where Support Vector Machine, a form of a high dimensional linear classifier, can be used to separate the data points [44]. Under certain mathematical conditions, we have a freedom to define the kernel function by giving the formula for each pair of observations.

In this section, we show how the Fisher information matrix defines a natural distance over a possibly multimodal representation of complex instances. Our goal is to define a unified kernel function with the following properties:

1. A single kernel should include all modalities to avoid the computational complexity of the multiple kernel learning problem and in particular the need for wrapper methods.
2. The kernel should be based on an underlying model that captures the connection and dependencies between the modalities or the multiple representations.
3. Data points should posses a generative model so that the Fisher information matrix can be used to define a mathematically justified optimal kernel.

3.1 Random Field representation

As the first step, we represent our data as a Random Field by assuming that the data instances are generated by defining their distances from certain selected instances S. Practically, we will select the training instances or, in case of too many of them, a subset of the training set but we may in fact use an arbitrary sample S.

We will consider our data points as random variables forming a Markov Random Field described by an undirected graph. For a target instance x, we define a generative model by a simple graph that has edges between x and each elements of the sample S.

We define a generative model of x based on its similarity or distance $\mathrm{dist}(x, s)$ to elements of sample S. In this random field, the factor graph is a star that consists of the pairs of x connected to the elements $s \in S$. By the Hammersley–Clifford theorem [40], the joint distribution of the generative model for X is a Gibbs distribution. Next we derive this distribution via an appropriate potential function.

3.2 The potential function

Given a Markov Random Field defined by a graph, a wide variety of proper potential functions can be used to define a Gibbs distribution. The weak but necessary restrictions are that the potential function has to be positive real valued, additive over the maximal cliques of the graph, and more probable configurations (specific sets of parameters) have to have lower potential.

Our first and least complex graph is a bipartite graph connecting only the actual observations and the finite set of previously known observations. For simplicity first we will discuss the single modality case. In this graph the maximal cliques are the pairs of the actual observation and the elements of the sample set, therefore our potential function can have a really simple form,

$$U(X \mid S, \theta = \{\alpha_i\}) = \sum_{i=1}^{|S|} \alpha_i \mathrm{dist}(x, s_i), \tag{1}$$

where θ is the hyperparameter and $s_i \in S$ is the ith sample.

For K modalities with different distance functions between the instances, the potential function has the form

$$U(x \mid S, \theta = \{\alpha_{ik}\}) = \sum_{i=1}^{|S|} \sum_{k=1}^{K} \alpha_{ik} \mathrm{dist}_k(x, s_i), \tag{2}$$

where K is the number of different distance functions and $\theta = \{\alpha_{ik}\}$ is the set of the parameters. For simplicity, from now on we omit S and use θ to denote the hyperparameters.

Given the potential function over the maximal cliques, by the Hammersley–Clifford theorem [40], the joint distribution of the generative model for X is a Gibbs distribution

$$p(X \mid \theta) = \mathrm{e}^{-U(X|\theta)}/Z(\theta) \tag{3}$$

where

$$Z(\theta) = \int_{X \in \mathcal{X}} \mathrm{e}^{-U(X|\theta)} \mathrm{d}X \tag{4}$$

is the expected value of the energy function over our generative model, a normalization term called the partition func-

tion. If the model parameters are previously determined, Z is a constant.

3.3 The Fisher Kernel

According to Jaakkola and Haussler [27], generative models have a natural kernel function based on the Fisher information matrix F.

The main innovation of Jaakkola and Haussler [28] is to obtain the kernel function *directly from a generative probability model* and therefore obtain a kernel quite closely related to the underlying model. They consider a parametric class of probability models $P(X|\theta)$ where $\theta \in \Theta \subseteq \mathbb{R}^l$ for some positive integer l.

Provided that the dependence on θ is sufficiently smooth, the collection of models with parameters from Θ can then be viewed as a (statistical) manifold M_Θ. M_Θ can be turned into a Riemannian manifold[1] [29] by giving a scalar product at the tangent space of each point. $P(X|\theta) \in M_\Theta$ via a positive semidefinite matrix $F(\theta)$, which varies smoothly with the base point θ. Such positive semidefinite matrices are provided by the Fisher information matrix

$$F(\theta) := \mathbf{E}(\nabla_\theta \log P(X|\theta) \nabla_\theta \log P(X|\theta)^T),$$

where the gradient vector $\nabla_\theta \log P(X|\theta)$ is

$$\nabla_\theta \log P(X|\theta) = \left(\frac{\partial}{\partial \theta_1} \log P(X|\theta), \ldots, .. \frac{\partial}{\partial \theta_l} \log P(X|\theta) \right),$$

and the expectation is taken over $P(X|\theta)$. In particular, if $P(X|\theta)$ is a probability density function, then the ij-th entry of $F(\theta)$ is

$$f_{ij} = \int_X P(X|\theta)(\frac{\partial}{\partial \theta_i} \log P(X|\theta))(\frac{\partial}{\partial \theta_j} \log P(X|\theta))dX.$$

In many cases the kernel can actually be viewed as an inner product:

$$K(X,Y) = \phi_X^T \phi_Y,$$

where the feature vectors $\phi_X, \phi_Y \in \mathbb{R}^k$ are obtained via a fixed, problem specific map $X \mapsto \phi_X$ which describes the examples X in terms of a real vector of length k.

The vector $G_X = \nabla_\theta \log P(X|\theta)$ is called the *Fisher score* of the example X. Now the mapping $X \mapsto \phi_X$ of examples to feature vectors can be $X \mapsto F^{-\frac{1}{2}} G_X$ (we suppressed here the dependence on θ), the *Fisher vector*. Thus, to capture the generative process, the gradient space of the model space M_Θ is used to derive a meaningful feature vector. The corresponding kernel function

$$K(X,Y) := G_X^T F^{-1} G_Y$$

is called the *Fisher kernel*.

An intuitive interpretation is that G_X gives the direction where the parameter vector θ should be changed to fit best the data X [36].

[1] A Riemannian manifold M is a smooth real manifold, where for each point $p \in M$ there is an inner product defined on the tangent space of p. This inner product varies smoothly with p. One can define the length of a tangent vector via this inner product on the tangent space. This makes possible to define the length of a curve $\gamma(t)$ on M by integrating the length of the tangent vector $\dot{\gamma}(t)$. This in turn allows to define a metric on M. The distance between two points Q and Q' is just the length of the shortest curve on M from. Q to Q'.

3.4 Fisher Kernel over Markov Random Fields

In this section we prove that the Fisher Information matrix assuming Gibbs distribution with potential function (1) is the variance matrix of the distances $\text{dist}_k(x, s_i)$ for $s \in S$, and therefore the Fisher kernel is the linear kernel over the normalized distances.

First, let us calculate the Fisher score based on our general generative model,

$$\begin{aligned} G_X^i &= \nabla_{\theta_i} \log p(X|\theta) \\ &= -\frac{\partial(U(X|\theta)}{\partial \theta_i} + \frac{1}{Z(\theta)} \int_{X \in \mathcal{X}} \mathrm{e}^{U(X|\theta)} \frac{\partial(U(X|\theta)}{\partial \theta_i} \mathrm{d}X. \end{aligned} \quad (5)$$

As we set our model θ fixed, $Z(\theta)$ is a constant and our formula can be simplified as

$$G_X^i = \mathbf{E}_\theta[\frac{\partial(U(X|\theta)}{\partial \theta_i}] - \frac{\partial(U(X|\theta)}{\partial \theta_i}. \quad (6)$$

The first part of the formula can be calculated from the observation X while the expected value (the mean of the gradient of the potential function) is hard to compute. Worth to mention, if there exists a probability density function $f(X \mid \theta)$ such that

$$U(X \mid \theta) = -\log f(X \mid \theta) \quad (7)$$

then the expected term of (6) is zero trivially. For a potential function as in equation (1), the Fisher score of X has a simple form,

$$G_X^i = \mathbf{E}_\theta[\text{dist}(\mathbf{x}, \mathbf{s_i})] - \text{dist}(x, s_i). \quad (8)$$

Before we move on to the analysis of the dimensionality, let us examine the computational properties of the Fisher information matrix.

3.5 Approximation of the Fisher Kernel over Gibbs distribution

The computational complexity of the Fisher information matrix is $\mathcal{O}(N|\theta|^2)$ where N is the size of the training set. The linearization of the Fisher kernel through Cholesky decomposition is also an expensive procedure depending only on the size of the parameter set.

To reduce the complexity to $\mathcal{O}(N|\theta|)$ we can approximate the Fisher information matrix with the diagonal as suggested in [27, 36].

Focusing on the diagonal of the Fisher information matrix, we get

$$\begin{aligned} f_{i,i} &= \mathbf{E}_\theta[\nabla_{\theta_i} \log p(X|\theta)^T \nabla_{\theta_i} \log p(X|\theta)] \\ &= \mathbf{E}_\theta[(\mathbf{E}_\theta[\frac{\partial(U(X|\theta)}{\partial \theta_i}] - \frac{\partial(U(X|\theta)}{\partial \theta_i})^2] \\ &= \int_{X \in \mathcal{X}} p(X \mid \theta)(\mathbf{E}_\theta[\frac{\partial(U(X|\theta)}{\partial \theta_i}] - \frac{\partial(U(X|\theta)}{\partial \theta_i})^2 \mathrm{d}X. \end{aligned} \quad (9)$$

For the potential function of equation (1), the diagonal of the Fisher kernel is the standard deviation of the distances from the samples and therefore the Fisher vector of X has the following form

$$\mathcal{G}_X^i = F_{ii}^{-\frac{1}{2}} G_X^i = \frac{\mathbf{E}_\theta[\text{dist}(x, s_i)] - \text{dist}(x, s_i)}{\mathbf{E}_\theta^{\frac{1}{2}}[(\mathbf{E}_\theta[\text{dist}(x, s_i)] - \text{dist}(x, s_i))^2]} \quad (10)$$

The above formula can be directly computed from the distance matrix of the sample S and the training and testing instances X. The dimensionality of the Fisher vector (the normalized Fisher score) is equal to the size of the parameter set of our joint distribution. In our case it depends only on the size of the sample S and the number of modalities (K), $dim_{Fisher} = K \cdot |S|$.

4. EXPERIMENTS

We performed experiments on publicly available real-world datasets from various domains. Next, we briefly describe the datasets, the underlying domains followed by the experimental protocol, results and discussion.

In all our experiments, we approximate the mean and variance of $\text{dist}_k(x, s_i)$ from the training data to compute the kernel as defined by equation (10). Since kernel methods are feasible for regression [38, 44], we also use the methods for predicting numerical values.

We used LibSVM [12] for classification problems and the Weka implementation of SMOReg [54][38] for regression.

Table 1: EEG prediction

Method	AUC	Gain(%)
DTW k-NN k=1	0.7534	+0.0
DTW k-NN k=100	0.7847	+0.0%
SimKer: 64xDTW |S|=100	0.8275	+5.4%
SimKer: MultiDTW |S|=|T|	0.8506	+8.4%

Table 2: Visual concept detection over the Yahoo! MIR Flickr dataset

Method	Mod.	MiAP	Gain(%)
ColHOG (CH)	Vis.	0.3670	
SimKer: Flickr tags (Sim.JS)	Text.	0.3015	
SimKer: CH + JS (Sim.JSCH)	Multi	0.4257	+2.0%
L.Comb: CH + Sim.JS	Multi	0.4170	+0.0%
L.Comb: Sim.JSCH + CH + Sim.JS	Multi	0.4467	+7.1%
SLWF by Liu (2014) [33]	Multi	0.4367	

Table 3: Quality prediction over the C3 dataset

Method	Mod.	MAE	RMSE	Gain(%)
BM25 SVM (BM)	Text.	0.6144	0.7915	+0.0%
C3 features GBT (GBT)	Netw.	1.3528	1.4961	
Lin. Comb.: BM + GBT	Multi	0.7459	0.8839	
SimKer: BM25	Text.	0.6196	0.8095	
SimKer: C3	Netw.	0.6900	0.8278	
SimKer: BM25 + C3	Multi	0.5891	0.7753	+4.2%

4.1 Time series classification

We performed experiments on the publicly available EEG dataset [58] from UCI machine learning repository[2] and the time series datasets from the UCR time series archive.[3]

For the classification of time-series, the k nearest-neighbor (k-NN) method using dynamic time warping (DTW) as distance measure was reported to be competitive, if not superior, to many state-of-the-art time-series classifiers, such as neural networks, hidden Markov models or support vector machines, see e.g. [15, 24, 55] and the references therein. Furthermore, Chen et al. [14] gave theoretical guarantees for the performance of nearest neighbor-like classifiers for time series. Therefore, we use k-NN with DTW as baseline.

EEG (electroencephalogram) is usually recorded on multiple channels, therefore, multimodality naturally arises with

[2]http://archive.ics.uci.edu/ml/datasets/EEG+Database
[3]www.cs.ucr.edu/~eamonn/time_series_data

Table 4: Quality prediction over the C3 dataset

Method	Mod	AUC	Gain(%)
tf SVM linear	Text.	0.6531	
tf SVM poly. d=2	Text.	0.6498	
tf SVM poly. d=3	Text.	0.6530	
tf.idf SVM linear	Text.	0.6496	
tf.idf SVM poly. d=2	Text.	0.6428	
tf.idf SVM poly. d=3	Text.	0.6464	
BM25 SVM linear (Lin)	Text.	0.6923	
BM25 SVM poly. d=2	Text.	0.6826	
BM25 SVM poly. d=3	Text.	0.6714	
C3 features LibFM	Netw.	0.6695	
C3 features GBT	Netw.	0.6688	
L.Comb.: Lin + LibFM	Multi	0.7100	
L.Comb.: Lin + GBT	Multi	0.7133	+0.0%
SimKer: tf JS (Sim.JS)	Text.	0.6978	
SimKer: BM25 L2 (Sim.BM)	Text.	0.7141	
SimKer: C3	Netw.	0.6571	
SimKer: BM+JS+C3 (Sim.All)	Multi	0.7363	+3.2%

Table 5: Web Spam detection over ClueWeb dataset

Method	Mod.	AUC	Gain(%)
BM25 SVM	Text.	0.8450	
Content features	Cont.	0.7882	
L.Comb.: BM + Cont.	Multi	0.8517	+0.0%
SimKer: BM25	Text.	0.8546	
SimKer: BM25 + Cont.	Multi	0.8622	+1.2%

Table 6: Classification of gene expression data

Method	AUC	Gain(%)
Linear SVM	0.9338	
Cosine SVM	0.9496	+0.0%
SimKer: cosine distance	0.9588	+0.9%

such data. Classification of EEG signals is one of the most prominent application domains in the light of ongoing American and European large scale research projects dedicated to study the brain and its disorders, such as the BRAIN Initiative [4] and the European Human Brain Project[5]. EEG is one of the most well-established techniques to capture the activity of the brain, it is widely used in research and clinical practice, see e.g. [2, 20, 42]. Paralyzed patients may benefit from EEG-controlled devices, such as spelling tools [8] or web browsers [5]. Furthermore, there were attempts to predict upcoming emergency braking based on EEG signals [26] which could result in reducing the braking distance of vehicles. A common feature of the aforementioned applications is that they involve classification of EEG signals.

The UCI EEG collection [58] contains in total 11028 EEG signals recorded from 122 persons. The total (decompressed) size of the data is several gigabytes which is roughly three orders of magnitude larger than the datasets from the UCR repository. Out of the 122 persons, there are 77 alcoholic patients and 45 healthy individuals. While capturing EEG, both alcoholic patients and healthy individuals were exposed to three different stimuli: subjects were shown either one picture or two different pictures or the same picture twice.

[4]http://en.wikipedia.org/wiki/BRAIN_Initiative
[5]https://www.humanbrainproject.eu

The dataset contains recordings for all the tree types of stimuli for all the subjects. Each signal was recorded using 64 electrodes at 256 Hz for 1 second. Therefore, each EEG signal is a 64-dimensional time series of length 256 in this collection. Multimodality, a core aspect of the proposed technique, naturally arises with multidimensional time series: each channel may correspond to a modality.

As a noise filter, a simple preprocessing step, we reduced the length of the signals from 256 to 64 by binning with a window size of four, i.e., we averaged four consecutive values of the signal.

In order to simulate the clinically relevant scenario in which the classifier is applied to the EEG of new patients, we randomly assign each person to either training or test split of the data and *all* the signals of the same person were either assigned to the training set or to the test set. In total, randomly selected 50 % of the all persons were assigned to the training set, while the remaining persons were assigned to the test set.

We performed two experiments on EEG data. In the first experiment, we randomly selected 100 signals as sample set S and calculated the DTW distances between these reference signals and other train and test signals for each channel *separately*. This experiment simulates application scenarios in which classification time is essential: in order to classify a new time series, we only need to calculate its distance to relatively few reference signals and use these distances as features in our approach. This allows quick and accurate classification of new signals. As the third row of Table 1 shows, our approach outperforms the baseline in terms of AUC.

In the second experiment, we used multivariate DTW as distance of two EEG signals. For a detailed description of multivariate DTW we refer to [9]. In this experiment, the distances from all the training signals were used as features in our approach, SimKer. While the DTW-calculations in this scenario require non-negligible computational effort, as Table 1 shows, this results in further improvements in terms of classification accuracy as measured by AUC.

Additionally, we performed experiments on the datasets of UCR time series archive which is one of the most frequently used benchmark in the time series literature. Note that the datasets in this collection are rather small, a few megabytes each, therefore, training advanced models on the datasets from the UCR collection is inherently difficult. Consequently, the advantage of complex models to simpler ones may not be pronounced on the UCR time series datasets, and we do not expect to observe substantial differences between different models on the UCR time series. In our approach, SimKer, we used DTW as distance measure and considered the distances from each training time series.

The results on the datasets of the UCR archive show that our approach clearly outperformed the baseline on some of the datasets of the archive, while the overall difference between the performance of our approach and the baseline was not found to be statistically significant using paired t-test at significance level of 0.05. We note that while we performed experiments on the data from the UCR time series archive, we considered only *one* modality (the DTW-distance of a time series x from the train time series), because no other modality was available for this data. Therefore, we could not exploit one of the major advantages of the proposed method, i.e., its ability to fuse several modalities.

4.2 Gene Expression

Proteins play essential role in almost all biological processes at the cellular level. Genes are particular subsequences of DNA that code for proteins. While each cell of the organism has the same DNA, the activation levels of genes may vary in different tissues: informally speaking, the expression level of a gene means how frequently the corresponding DNA fragment is transcribed to RNA and translated to proteins. Various tissues are characterized by different gene expression patterns, furthermore, diseases such as cancer may be associated with characteristic gene expression patterns. Therefore, classification of gene expression data may contribute to diagnosis of various types of cancer such as colon cancer, lymphoma, lung cancer or subtypes of breast cancer [32]. In this paper, we used publicly available gene expression data of breast cancer tissues, colon cancer tissues, and lung cancer tissues, see [32] and the references thereoin for details. In these datasets, the expression levels of 7650, 6500 and 12,600 genes have been measured for 95, 62 and 203 patients in the breast cancer, colon cancer and lung cancer datasets respectively.

Similarly to [32], we performed experiments according to the 5-fold crossvalidation protocol. As baselines, we used SVMs, because SVMs were reported to perform excellently on these datasets.

Table 6 summarizes our results: we report AUC averaged over all the three datasets for SimKer and SVMs with linear and cosine kernel. The results show that SimKer outperforms both types of SVMs.

4.3 Web Spam detection over ClueWeb09

The first results on automatic Web quality classification focus on Web spam [11]. In this section, we show experiments over the Waterloo Spam Rankings [16] of the ClueWeb09 corpus.

Our baseline classification procedures are collected in [48] by analyzing the results of the Web Spam Challenges and the ECML/PKDD Discovery Challenge 2010. As our main conclusion, Web spam can be classified purely based on the terms used. Over different Web spam and quality corpora [22], the bag-of-words classifiers based on the top few 10,000 terms performed best and significantly improved the traditional Web spam features [11]. SVM based content classification was first used in [1]. In our earlier result, we use libSVM [12] with several kernels and apply late fusion as described in [48]. We improve over this later result by using the Fisher kernel next.

Our most important feature set is the bag of words representation of the text over the Web host. Let there be H hosts consisting of an average $\overline{\ell}$ terms. Given a term t of frequency f over a given host that contains ℓ terms, we used the BM25 term weightning scheme, where the weight of t in the host becomes

$$\log \frac{H-h+0.5}{h+0.5} \cdot \frac{f(k+1)}{f+k(1-b+b\cdot\frac{\ell}{\overline{\ell}})}. \qquad (11)$$

Low k means very quick saturation of the term frequency function while large b downweights content from very large Web hosts.

In addition, we use the public feature set [10] that includes the following values computed for the home page, page with the maximum pagerank and average over the entire host:

1. Number of words in the page, title;
2. Average word length, average word trigram likelihood;
3. Compression rate, entropy;
4. Fraction of anchor text, visible text;
5. Corpus and query precision and recall.

Here feature classes 1–4 can be normalized by using the average and standard deviation values over the two collections while class 4 is likely domain and language independent.

Corpus precision and recall are defined over the k most frequent words in the dataset, excluding stopwords. Corpus precision is the fraction of words in a page that appear in the set of popular terms while corpus recall is the fraction of popular terms that appear in the page. This class of features is language independent but rely on different lists of most frequent terms for the two data sets.

Results for spam detection in Table 5 show 1.2% improvement for the multimodal Similarity kernel over the linear combination of the predictions of the BM25 based SVM and the content feature based SVM.

4.4 Web credibility classification

Mining opinion from the Web and assessing its quality and credibility became a well-studied area [21]. Classifying various aspects of quality was introduced as part of the ECML/PKDD Discovery Challenge 2010 tasks [48] and among others, Microsoft created a reference data set [46].

Recent results on Web credibility assessment [34] use content quality and appearance features combined with social and general popularity and linkage. After feature selection, they use 10 features of content and 12 of popularity by standard machine learning methods of the scikit-learn toolkit.

In this section we show the performance of the Fisher kernel for the WebQuality 2015 Data Challenge by comparing prediction methods for the C3 data set. The data set was created in the Reconcile[6] project and contains 22325 Web page evaluations in five dimensions (credibility, presentation, knowledge, intentions, completeness) of 5704 pages given by 2499 people. The mTurk platform were used for collecting evaluations. Ratings are similar to the dataset built by Microsoft for assessing Web credibility [46], on a scale of four values 0-4, with 5 indicating no rating. Since multiple values may be assigned to the same aspect of a page, we simply average the human evaluations per page. We may also consider binary classification problems by assigning 1 for above 2.5 and 0 for below 2.5.

While we are aware of no other results over the C3 data set, we collect reference methods from Web credibility research results. Existing results fall in four categories: Bag of Words; language statistical, syntactic, semantic features; numeric indicators of quality such as social media activity; and assessor-page based collaborative filtering. User and page-based collaborative filtering is suggested in [35] in combination with search engine rankings. Social media and network based features appear already for Web spam [25, 11]. Content statistics as a concise summary that may replace the actual terms in the document were introduced first in the Web spam research [11]. The C3 data set includes content quality and appearance features described among others in [34].

[6] `http://reconcile.pjwstk.edu.pl/`

In order to perform text classification, we crawled the pages listed in the C3 data set. By using the bag of words representation of the Web page content, our goal is to combine all above methods with known and new kernel based text classifiers.

Our classifier ensemble consists of the following components:

- Gradient Boosted Trees and recommenders
- Standard text classifiers
- Similarity kernel based SVM using not only the text but also the C3 attributes.

In our experiments the Bag of words models contain the top $30k$ term frequencies after stemming. Besides BM25 (see Section 4.3), we experimented with two additional term frequency normalization schemes:

- Term frequency (tf): simply f, for all terms in the documents of H.
- Term frequency times inverse document frequency (tf.idf):

$$\log \frac{H - h + 0.5}{h + 0.5} \cdot f. \tag{12}$$

One of the main questions is how to select proper distance measures over the bag of words and C3 features. In addition to the linear metric over the C3 attributes and the L2 normalized bag of words representations (tf, tf.idf and BM25), we apply Jensen-Shannon divergence (JS) over the L1 normalized term distributions according to our previous results [48, 18].

Our most complex Fisher kernel (Sim.All) is based on three representations: Jensen-Shannon divergence over raw term distribution, Euclidean distance over L2 normalized BM25, Euclidean distance over scaled site features.

According the results in in Table 4 the use of Fisher kernel over the term frequency based Jensen-Shannon divergence (Sim.JS) already reaches accuracy of the best non Fisher method with a single modality (Lin, linear SVM over the BM25 features). Out of the non Fisher methods using only the C3 attributes the LibFM and the Gradient Boosted Tree (GBT) perform very similar. The ensemble of GBT and the linear SVM over BM25 performs 0.713 in AUC, achieving the accuracy of the best Fisher kernel with only one distance (Sim.BM).

The best method (Sim.All) outperforms the best non Fisher method (Linear combination of Lin and GBT) by 3.2% on average in AUC. The largest difference is 7.2% by classifying "knowledge". Similarity kernel performs similarly for regression (Table 3). We measured 4.2% improvement in MAE (Mean Absolute Error) and 2.1% in RMSE (Root Mean Squared Error) over the baseline method.

4.5 Visual concept detection: Yahoo! MIR Flickr dataset

Images are rarely being present alone, usually we can extract some content related textual or other non-visual information such as geo-location or date from their context. Besides non visual meta features we can think of any visual representation as an individual modality. Altogether we can easily define a set of very diverse distance functions over images.

Vast amount of tagged images is available over photo sharing services or even the public Web. In our experiments we

used the Yahoo! MIR Flickr dataset containing $15k$ images as the training set and $10k$ images as a test set [51]. The dataset was used for various challenges such as ImageCLEF 2012 Photo Annotation task [51] and in recent articles [33][7][50]. The aim is to detect the presence of 94 categories (a wide variety of concepts not limited to objects, e.g. daylight, indoor, underwater or citylife) in terms of their visual and textual features.

Among large number of Bag of Visual Words models (super vector [59], kernel codebook [52], locality-constrained [53] to name a few), Gaussian Mixture based Fisher encoding [37] appears best out of BoVW models by the evaluation work of [13][43], hence we choose the same method. The Fisher metric over Gaussian mixtures is a well-known method to measure the distance between two images based on their visual content [36, 13, 43]. The model extracts a large amount of local descriptors over various parts of the image. The Gaussian Mixture model describes the set of descriptors of the image assuming naive independence between the descriptors. In our experiments we calculated grayscale HOG (Histogram of Oriented Gradients [17]) and RGB color moments over a dense grid and multiple scales using four different macroblock sizes (24x24, 32x32, 48x48 and 64x64 pixels per block). Both descriptors were L2 normalized. To reduce the dimension of the descriptors we transformed the vectors by Principal Component Analysis (PCA). The procedure resulted approximately $140k$ descriptors per image. The final visual Fisher vectors with 512 Gaussians were calculated over the descriptors per image. Moreover we splitted the images into three parts according to Lazebnik et al. [41] increasing the number poolings per image.

Additionally, we computed Jensen-Shannon divergence of the images based on their Flickr tags. As a baseline, we combined linearly the predictions of the linear SVM over the Gaussian Mixture based visual Fisher kernel and the Similarity kernel of the Jensen-Shannon divergence over the Flickr tags. The multimodal Similarity kernel (JSCH) outperforms the baseline by 2% (see Table 2) in MiAP (Mean interpolated Average Precision, the metric at the task [51]). Our best method, surprassing the baseline by 7.1%, is a linear combination of the predictions using the visual Fisher kernel and the Similarity kernels, both textual and multimodal.

In comparison to recent results, our method outperforms the Selective Weighted Late Fusion (Liu et al. [33]) by 2.28%, the best result published to our knowledge over the MIR Flickr dataset.

5. CONCLUSIONS

From a generative model based on instance similarities, we derived a "similarity" kernel applicable for SVM classification and regression. The method is capable of defining a single unified kernel even in the case of rich data types, including multimodal or multiple time series data. The parameters of the kernel are directly computable from the data and hence we may avoid the high computational costs of multiple kernel learning and in particular the need for wrapper methods.

We evaluated our methods on a variety of publicly available real data sets, including multi-channel EEG, univariate time series, gene expression data, Web spam and credibility as well as image content with text annotation. Besides the presence of multiple modalities, complexity of classification and regression tasks in the aforementioned domains arise from various additional sources, such as high dimensionality (compared to the number of available instances), interdependence between attributes, presence of noise and uncertainty. Our experiments show that the proposed approach is able to successfully solve the underlying machine learning tasks, even under the presence of such additional domain and data complexity.

In particular, on all the aforementioned data sets, our method reaches and in many cases improves over the state-of-the-art. Hence we conclude generative models based on instance similarities with multiple modes is a generally applicable model for classification and regression tasks ranging over various domains, including but not limited to the ones presented in this paper.

6. ACKNOWLEDGEMENTS

This paper was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. Research partially performed within the framework of the grant of the Hungarian Scientific Research Fund (grant No. OTKA 111710).

The publication was supported in part by the EC FET Open project "New tools and algorithms for directed network analysis" (NADINE No 288956), by the Momentum Grant of the Hungarian Academy of Sciences, by OTKA NK 105645, the KTIA_AIK_12-1-2013-0037 and the PIAC_13-1-2013-0197 projects. The projects are supported by Hungarian Government, managed by the National Development Agency, and financed by the Research and Technology Innovation Fund.

7. REFERENCES

[1] Jacob Abernethy, Olivier Chapelle, and Carlos Castillo. WITCH: A New Approach to Web Spam Detection. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2008.

[2] Jessica Askamp and Michel J.A.M. van Putten. Diagnostic decision-making after a first and recurrent seizure in adults. *Seizure*, 22(7):507 – 511, 2013.

[3] Francis R Bach, Gert RG Lanckriet, and Michael I Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the twenty-first international conference on Machine learning*, page 6. ACM, 2004.

[4] Asa Ben-Hur and William Stafford Noble. Kernel methods for predicting protein–protein interactions. *Bioinformatics*, 21(suppl 1):i38–i46, 2005.

[5] Michael Bensch, Ahmed A Karim, Jürgen Mellinger, Thilo Hinterberger, Michael Tangermann, Martin Bogdan, Wolfgang Rosenstiel, and Niels Birbaumer. Nessi: an EEG-controlled web browser for severely paralyzed patients. *Computational intelligence and neuroscience*, 2007.

[6] Mikhail Bilenko and Raymond J Mooney. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 39–48. ACM, 2003.

[7] Alexander Binder, Wojciech Samek, Klaus-Robert Müller, and Motoaki Kawanabe. Enhanced

representation and multi-task learning for image annotation. *Computer Vision and Image Understanding*, 117(5):466–478, 2013.

[8] Niels Birbaumer, Nimr Ghanayim, Thilo Hinterberger, Iver Iversen, Boris Kotchoubey, Andrea Kübler, Juri Perelmouter, Edward Taub, and Herta Flor. A spelling device for the paralysed. *Nature*, 398(6725):297–298, 1999.

[9] Krisztian Antal Buza. *Fusion methods for time-series classification*. Peter Lang Verlag, 2011.

[10] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: web spam detection using the web topology. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 423–430, 2007.

[11] Carlos Castillo, Debora Donato, Luca Becchetti, Paolo Boldi, Stefano Leonardi, Massimo Santini, and Sebastiano Vigna. A reference collection for web spam. *SIGIR Forum*, 40(2):11–24, December 2006.

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

[13] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*, 2011.

[14] George H. Chen, Stanislav Nikolov, and Devavrat Shah. A latent source model for nonparametric time series classification. In *Advances in Neural Information Processing Systems 26*, pages 1088–1096. 2013.

[15] Yanping Chen, Bing Hu, Eamonn Keogh, and Gustavo EAPA Batista. Dtw-d: time series semi-supervised learning from a single example. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 383–391. ACM, 2013.

[16] G.V. Cormack, M.D. Smucker, and C.L.A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Information retrieval*, 14(5):441–465, 2011.

[17] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition (CVPR), 2005 IEEE Conference on*, 2005.

[18] B. Daróczy, A. Benczúr, and R. Pethes. Sztaki at imageclef 2011. *Working Notes of CLEF 2011*, 2011.

[19] Bálint Daróczy, Dávid Siklósi, and András A Benczúr. Dms-sztaki@ imageclef 2012 photo annotation. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[20] J. Dauwels, F. Vialatte, T. Musha, and A. Cichocki. A comparative study of synchrony measures for the early diagnosis of alzheimer's disease based on eeg. *NeuroImage*, 49(1):668 – 693, 2010.

[21] K. Dave, S. Lawrence, and D.M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM, 2003.

[22] Miklós Erdélyi, András Garzó, and András A. Benczúr. Web spam classification: a few features worth more. In *Joint WICOW/AIRWeb Workshop on Web Quality (WebQuality 2011) In conjunction with the 20th International World Wide Web Conference in Hyderabad, India.* ACM Press, 2011.

[23] Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011.

[24] Steinn Gudmundsson, Thomas Philip Runarsson, and Sven Sigurdsson. Support vector machines and dynamic time warping for time series. In *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 2772–2776. IEEE, 2008.

[25] Zoltán Gyöngyi and Hector Garcia-Molina. Spam: It's not just for inboxes anymore. *IEEE Computer Magazine*, 38(10):28–34, October 2005.

[26] Stefan Haufe, Matthias S Treder, Manfred F Gugler, Max Sagebaum, Gabriel Curio, and Benjamin Blankertz. Eeg potentials predict upcoming emergency brakings during simulated driving. *Journal of neural engineering*, 8(5):056001, 2011.

[27] Tommi S Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, pages 487–493, 1999.

[28] Tommi S Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, pages 487–493, 1999.

[29] Jürgen Jost. *Riemannian geometry and geometric analysis*. Springer, 2011.

[30] Gert RG Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I Jordan. Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research*, 5:27–72, 2004.

[31] Christina S Leslie, Eleazar Eskin, and William Stafford Noble. The spectrum kernel: A string kernel for svm protein classification. In *Pacific symposium on biocomputing*, volume 7, pages 566–575, 2002.

[32] Wei-Jiun Lin and James J Chen. Class-imbalanced classifiers for high-dimensional data. *Briefings in bioinformatics*, 14(1):13–26, 2013.

[33] Ningning Liu, Emmanuel Dellandrea, Bruno Tellez, and Liming Chen. A selective weighted late fusion for visual concept recognition. In *Fusion in Computer Vision*, pages 1–28. Springer, 2014.

[34] Alexandra Olteanu, Stanislav Peshterliev, Xin Liu, and Karl Aberer. Web credibility: Features exploration and credibility prediction. In *Advances in Information Retrieval*, pages 557–568. Springer, 2013.

[35] Thanasis G Papaioannou, Jean-Eudes Ranvier, Alexandra Olteanu, and Karl Aberer. A decentralized recommender system for effective web credibility assessment. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 704–713. ACM, 2012.

[36] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07*, pages 1–8, 2007.

[37] Florent Perronnin, Jorge Sánchez, and Thomas

Mensink. Improving the fisher kernel for large-scale image classification. In *Computer Vision–ECCV 2010*, pages 143–156. Springer, 2010.

[38] John C. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, ADVANCES IN KERNEL METHODS - SUPPORT VECTOR LEARNING, 1998.

[39] Alain Rakotomamonjy, Francis Bach, Stephane Canu, and Yves Grandvalet. simplemkl. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

[40] Brian D Ripley and Frank P Kelly. Markov point processes. *Journal of the London Mathematical Society*, 2(1):188–192, 1977.

[41] C. Schmid S. Lazebnik and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New York, June 2006*, 2006.

[42] Malihe Sabeti, Serajeddin Katebi, and Reza Boostani. Entropy and complexity measures for eeg signal classification of schizophrenic and control participants. *Artificial Intelligence in Medicine*, 47(3):263 – 274, 2009.

[43] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245, 2013.

[44] Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola, editors. *Advances in kernel methods: support vector learning*. MIT Press, Cambridge, MA, USA, 1999.

[45] Bernhard Schölkopf, Koji Tsuda, and Jean-Philippe Vert. *Kernel methods in computational biology*. MIT press, 2004.

[46] Julia Schwarz and Meredith Morris. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1245–1254. ACM, 2011.

[47] John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[48] D. Siklósi, B. Daróczy, and A.A. Benczúr. Content-based trust and bias classification via biclustering. In *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, pages 41–47. ACM, 2012.

[49] Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. Large scale multiple kernel learning. *The Journal of Machine Learning Research*, 7:1531–1565, 2006.

[50] B Thomee, M Huiskes, and M S. Lew. Special issue on visual concept detection in the mirflickr/imageclef benchmark. *Computer Vision and Image Understanding*, 117:451–452, 2013.

[51] B. Thomee and A. Popescu. Overview of the imageclef 2012 flickr photo annotation and retrieval task. *Working Notes of CLEF 2012, Rome, Italy*, 2012, 2012.

[52] Jan C van Gemert, Jan-Mark Geusebroek, Cor J Veenman, and Arnold WM Smeulders. Kernel codebooks for scene categorization. In *Computer Vision–ECCV 2008*, pages 696–709. Springer, 2008.

[53] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3360–3367. IEEE, 2010.

[54] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, second edition, June 2005.

[55] Xiaopeng Xi, Eamonn Keogh, Christian Shelton, Li Wei, and Chotirat Ann Ratanamahatana. Fast time series classification using numerosity reduction. In *Proceedings of the 23rd international conference on Machine learning*, pages 1033–1040. ACM, 2006.

[56] Guangnan Ye, Dong Liu, I-Hong Jhuo, and Shih-Fu Chang. Robust late fusion with rank minimization. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3021–3028. IEEE, 2012.

[57] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. *The Journal of Machine Learning Research*, 3:1083–1106, 2003.

[58] Xiao Lei Zhang, Henri Begleiter, Bernice Porjesz, Wenyu Wang, and Ann Litke. Event related potentials during object recognition tasks. *Brain Research Bulletin*, 38(6):531–538, 1995.

[59] Xi Zhou, Kai Yu, Tong Zhang, and Thomas S. Huang. Image classification using super-vector coding of local image descriptors. In *Proceedings of the 11th European conference on Computer vision: Part V*, ECCV'10, pages 141–154, Berlin, Heidelberg, 2010. Springer-Verlag.