様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　５　月　２０　日現在

機関番号：　２１６０２

研究種目：　基盤研究(C)　(一般)

研究期間：　2015～2017

課題番号：　１５Ｋ００２４３

研究課題名（和文）Speech based emotional and depressive mental state prediction using Gaussian Process state-space models

研究課題名（英文）Speech based emotional and depressive mental state prediction using Gaussian Process state-space models

研究代表者

ＭＡＲＫＯＶ　Ｋ（Markov, Konstantin）

会津大学・コンピュータ理工学部・上級准教授

研究者番号：８０３９４９９８

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：人の感情状態を推定することは、医療、社会的相互作用、サービス産業におけるアプリケーションの重要な課題です。一方、人格の認識は、感情推定をその構成要素の1つとして含む、さらに挑戦的な課題である。この研究では、感情認識のためのいくつかのシステムを構築するために、信号処理および機械学習分野における最新の開発および技術を使用した。使用される方法および技術には、ガウスプロセス、非線形状態空間モデルおよび深層ニューラルネットワークが含まれる。私たちの発見は、IEEE Accessジャーナルといくつかの国際会議に掲載されています。

研究成果の概要（英文）：Estimating the emotional state of a person is an important task with applications in medicine, social interaction as well as in services industry. On the other hand, recognition of the person personality is even more challenging task which includes emotion estimation as one of its components. In this research, we used latest developments and technologies in signal processing and machine learning fields to build several systems for emotion recognition from speech and music as well as text based personality recognition system. The used methods and technologies include Gaussian Processes, non-linear state-space models and deep neural networks. During the last year of this research, we focused on personality recognition task and built a system based on deep neural networks capable of recognizing the five personality traits with high accuracy. Our findings are published in the IEEE Access journal and several international conferences.

研究分野： Signal Processing, Machine Learning

キーワード： Speech emotion　Neural Networks　Gaussian Process　Personality Recogniition　Music emotion

１．研究開始当初の背景

Emotions perceived by humans from speech as well as from music have common psychological ground and studies in this area can be conducted using both speech and music as emotion information source. Estimating the emotional state of a person is an important task with applications in medicine, social interaction as well as in services industry. On the other hand, recognition of the person personality is even more challenging task which includes emotion estimation as one of its components.

２．研究の目的

The purpose of this project is to develop new technology for estimation of the emotional state of a person through the specific changes in the manner and the way of speaking. Closely related is the task of estimating emotions induced by music which is useful in music recommendation systems. On the other hand, recognition of the person personality is even more challenging task which includes emotion estimation as one of its components. Development of state-of-the-art music emotion recognition system as well personality recognition system based on the latest achievements in signal processing and machine learning is the other purpose of this research.

３．研究の方法

In this research, we have used several state-of-the-art modeling frameworks such as Gaussian Processes, Non-linear state-space models and Deep Neural Networks. Gaussian Processes are becoming more and more popular in the Machine Learning community for their ability to learn highly non-linear mappings between two continuous data spaces. Previously, we have successfully applied GPs for music genre classification task us to use GPs for emotion estimation. Many previous studies have focused on Support Vector regression (SVR) since in most cases it gives superior performance. We compare GP regression with SVR and show that in certain cases GPR can significantly outperform SVR. In addition, GPR produces probabilistic predictions, i.e. it outputs a Gaussian distribution with mean which corresponds to the most probable target value and variance which shows the certainty of the prediction. As in the case of SVR, GPR also uses kernels, but in contrast, it allows kernels parameters to be learned from the training data. Our approach is to consider emotion trajectories as time series and apply methods from time series analysis. One widely used method is Bayesian filtering based on state-space models (SSMs). A classic example is the Kalman filter. It has been successfully used for temporal music emotion recognition. However, the Kalman filter is a linear system and has its limitations. There exist non-linear SSMs such as the Extended Kalman filter (EKF) and Unscented Kalman filter (UKF), but they put certain constraints on the SSM state and measurement functions. A better solution is to use Gaussian Processes (GPs) which are non-linear, non-parametric models. They have been successfully applied in various tasks including speech and music processing. A number of GP based state-space models (GP-SSM) have been proposed recently. GP-BayesFilters use GPs as non-linear functions and derive GP-Particle filter, GP-EKF, and GP-UKF algorithms using Monte Carlo sampling. An analytic filtering approximation algorithm is presented, but lacks an analytic approach to GP-SSM parameter learning. An attempt to derive such an algorithm is done, which, however, has some stability problems. A Particle Markov Chain Monte Carlo (PMCMC) training method is described], but the MC based learning is notoriously slow. When we apply a SSM for continuous speech emotion recognition, states x_t would represent the unknown affect vector, i.e., Arousal-Valence-Dominance values, and y_t would correspond to feature vectors extracted from the speech signal. When observations of the state variable are available during training, $f()$ and $g()$ can be learned independently which makes the SSM parameter estimation simpler. Having non-linear functions $f()$ and $g()$ would greatly increase the expressiveness of the state-space model, but introduces two problems - what kind of non-linearity is suitable for the task at hand and how to estimate the parameters. Gaussian Processes allow eliminating the first problem and, when state observations are available, provide solution to the second. In GP inference, the non-linear function is marginalized out and there is no need to define it. The GP kernel function parameters can be learned using approximations and gradient descent methods. However, filtering with SSM when $f()$ and $g()$ are described by GPs is not straightforward. There are just a few studies on this problem and no common and efficient algorithm exists yet. In our experiments, we adopted new solution. It is based on analytic moment matching to derive Gaussian approximation to the filtering distribution. In addition, we implemented a Particle filter based approximation. Automatic recognition of person's personality from his/her social network activities allows to make predictions about preferences across contexts and environments and has many important practical applications, such as products, jobs, or services recommendation, word polarity disambiguation, mental health diagnosis, etc. Many approaches have been proposed to automatically infer users' personality from the content they generate in social networks. However, the performance of these approaches depends heavily on the data representation which often is based on hard-coded prior knowledge. Recently, deep learning approaches have obtained very high performance across many different natural language processing (NLP) tasks. Unlike traditional methods, deep learning approaches can learn suitable representation automatically. In this project, we implemented several deep learning algorithms including fully-connected neural networks (FC), convolutional neural networks (CNN) and recurrent neural networks (RNN) in our personality recognition system and evaluated it on the task from the "Workshop on Computational Personality Recognition (Shared Task)". Currently, word embedding has become a standard component for the DNN based natural language processing. It converts the one-hot representation of the word to a distributed representation, which has many benefits and allows to map words with similar meaning to similar values: the learning of one word can indirectly help the learning of the other words with similar meaning. This is especially helpful for tasks with small training data. In order to utilize the statistical knowledge of the text, we pre-train the word embedding matrix with the text data using the skip-gram method. We didn't use Google pre-trained word2vec because the statuses contain internet-slang, emoticons (e.g., :-D), acronyms (e.g., BRB-be right back) and various shorthand notations, which carry rich information about personality, but are not included in the Google model.

４．研究成果

The GP and SVM have many common characteristics. They are both non-parametric, kernel based models, and their implementation and usage as regressors or binary classifiers are the same. However, GP are probabilistic Bayesian predictors which in contrast to SVM produce Gaussian distributions as their output. Another advantage is the possibility of parameter learning from the training data. On the other hand, SVM provide sparse solution, i.e. only "support" vectors are used for the inference, which can be a plus when working with large amount of data.

For the speech emotion recognition task, we have developed and investigated a dynamic speech emotion recognition system using two different state-space models such as linear Kalman filter and a novel non-linear, non-parametric Gaussian Processes-based SSM. Kalman filters are widely used and well known SSM. On the other hand, the GP based models are new and there are few studies focusing on learning and inference algorithms for them. We were able to simplify the GP-SSM learning by utilizing the AVEC 2014 database which provides ground truth labels for the latent affect states. For the filtering and smoothing, however, there is no common and efficient algorithm. We compared the performance of a recently proposed analytic approximation based algorithm and a GP based Particle filter in terms of Pearson correlation coefficient and root mean square error with respect to the conventional Kalman filter. Both GP filtering algorithms showed about two times better results when the same feature vectors are used. A disadvantage of the GP-SSMs, however, is their memory and computational complexity which is much higher.

In the personality recognition task, we applied deep learning approaches including convolutional neural networks and recurrent neural networks. The results showed that FC models with only text are worse than the ones with author information. DNN with both text and author information achieves the best results. CNN, RNN and FC are able to automatically extract useful features for personality recognition and the best result of 60.0±6.5% F1 score was obtained using CNN with average pooling. We found that bi-gram, tri-gram and recurrent architecture didn't get better results. It may indicate that words chosen by the author tell more about author's personality than the meaning author expresses. We also noticed that applying regularization barely restrains the overfitting, the network learns the patterns that only exist in the training set. This may be improved by collecting more data or by transforming the text into a representation which is stable for personality by external knowledge.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　1　件）

1. K.Markov, T.Matsui, “Music Genre and Emotion Recognition Using Gaussian Processes", Access, IEEE, v.2, pp.688-697, 2014.

〔学会発表〕（計　5　件）

1. J.Yu, K.Markov, “Deep learning based personality recognition from Facebook status updates”, in Proc. IEEE 8th Int. Conf. on Awareness Science and Technology, iCAST 2017, pp.383-388, 2017.
2. K.Markov, T.Matsui, "Robust Speech Recognition using Generalized Distillation Framework", In Proc. Interspeech, pp.2364-2368, Sep 2016.
3. J.Yu, K.Markov, T.Matsui, "Articulatory and Spectrum Features Integration using Generalized Distillation Framework", IEEE Int. Workshop on Machine Learning for Signal Processing, Sep. 2016.
4. K.Markov, T.Matsui, F.Septier, G.Peters, "Dynamic Speech Emotion Recognition with State-Space Models", In Proc. European Signal Processing Conference, pp.2122-2126, Sep. 2015.
5. M.Soleymani, A.Aljanaki, Y.Yang, M.Caro, F.Eyben, K.Markov, B.Schuller, R.Veltkamp, F.Weninger, F.Wiering, "Emotional Analysis of Music: A Comparison of Methods", In ACM International Conference on Multimedia, pp.1161-1164, Nov. 2014

〔図書〕（計　　件）

〔産業財産権〕

○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
University of Aizu（　　　）

研究者番号：80394998

(2)研究分担者
Institute of Statistical Mathematics（　　　）

研究者番号：10370090

(3)連携研究者
（　　　）

研究者番号：

(4)研究協力者
（　　　）