INTERNATIONAL JOURNAL OF TECHO-ENGINEERING

ISSN : 2057-5688

TUMK-ELM A FAST UNSUPERVISED HETEROGENEOUS DATA LEARNING APPROACH

BANDI SAIRAM
PG Scholar, Department of Computer Science,
SVKP & Dr K S Raju Arts & Science College(A),
Penugonda, W.G.Dt., A.P, India
bandisr3@gmail.com

PADALA SRINIVASA REDDY
Associate Professor in Computer Science,
SVKP & Dr K S Raju Arts & Science College(A),
Penugonda, W.G.Dt., A.P, India
psreddy1036@gmail.com

Abstract

Advanced unsupervised learning techniques are an emerging challenge in the big data era due to the increasing requirements of extracting knowledge from a large amount of unlabeled heterogeneous data. Recently, many efforts of unsupervised learning have been done to effectively capture information from heterogeneous data. However, most of them are with huge time consumption, which obstructs their further application in the big data analytics scenarios, where an enormous amount of heterogeneous data are provided but real-time learning are strongly demanded. In this paper, we address this problem by proposing a fast unsupervised heterogeneous data learning algorithm, namely two-stage unsupervised multiple kernel extreme learning machine (TUMK-ELM). TUMK-ELM alternatively extracts information from multiple sources and learns the heterogeneous data representation with closed-form solutions, which enables its extremely fast speed. As justified by theoretical evidence, TUMK-ELM has low computational complexity at each stage, and the iteration of its two stages can be converged within finite steps. As experimentally demonstrated on 13 real-life data sets, TUMK-ELM gains a large efficiency improvement compared with three state-of-the-art unsupervised heterogeneous data learning methods (up to 140 000 times) while it achieves a comparable performance in terms of effectiveness.

keywords : Kernel, Task analysis, Unsupervised learning, Big Data, Data mining, Machine learning

1.INTRODUCTION

1.1 Introduction:

In most real-world data analytics problems, a huge amount of data are

ISSN : 2057-5688

collected from multiple sources without label information, which is often with different types, structures, and distributions, namely heterogeneous data. For example, in a sentiment analysis task, the data may contain texts, images, and videos from Twitters, Facebook, and YouTube. For extracting knowledge from such big unlabeled heterogeneous data, advanced unsupervised learning techniques are required to (1) have a large model capacity/complexity, (2) have the ability to integrating information from multiple sources and (3) have a high learning speed. Recently, many researchers enhance model capacity by combining unsupervised learning with deep learning to propose deep unsupervised learning models. These models inherit the powerful model capacity from deep neural networks that can reveal highly complex patterns and extremely nonlinear relations. However, most of them fail to learn from multiple sources. They are challenged by types, relations and distributions of the heterogeneous data because of the deep neural networks they used.

Without strong supervised information, the deep neural networks may arbitrarily fit complex heterogeneous data that leads to meaningless solutions. One promising way to reveal information from multiple sources is using multiple kernel learning (MKL, for short). MKL first adopts multiple kernels to capture heterogeneous data characteristics from different sources. It then learns optimal combination coefficients for these kernels guided by a specific learning task. In this way, MKL can effectively capture different complex distributions by different kernels, and reveal the relations between these different distributions by the kernel combination coefficients. Despite the advantages of MKL, it requires supervised label information to learn the optimal kernel combination coefficients. However, label information is often not available or very costly in real big data analytics task, which limits the application of MKL.

More recently, unsupervised MKL has been studied to tackle the heterogeneous data learning without supervised labels. Similar to MKL, unsupervised MKL also uses multiple kernels to distill information from various sources. To enable the learning without supervised labels, it introduces a kernel-based unsupervised learning objective, e.g. kernel k-means, to learn the optimal kernel combination coefficients.

ISSN : 2057-5688

Although unsupervised MKL achieves remarkable performance in unsupervised heterogeneous data learning, most of the current unsupervised MKL methods are with a slow learning speed.

The slow learning speed is mainly caused by the iterative numerical solution, which is adopted by these methods for optimizing the kernel combination coefficients. It does not satisfy the requirements of (1) handling a large amount of data and (2) real-time learning. To address the above issues, we here propose a fast unsupervised heterogeneous data learning approach, namely Two-stage Unsupervised Multiple Kernel Extreme Learning Machine (TUMK-ELM, for short). TUMK-ELM iteratively extracts information from multiple sources and learns the heterogeneous data representation with closed-form solutions. It adopts multiple kernels to capture information in heterogeneous data and learns an optimal kernel for heterogeneous data representation.

Different from current unsupervised multiple kernel learning methods, it seamlessly integrates a much more efficient kernel combination coefficients optimization method with an effective unsupervised learning objective that simultaneously guarantees a fast learning speed and a high learning quality. Specifically, TUMK-ELM uses the kernel k-means objective function to guide the unsupervised learning process and adopts the distance-based multiple kernel extreme learning machine (DBMK-ELM, for short) to learn the kernel combination coefficients. TUMK-ELM can be split into two iterative stages.

At the first stage, TUMK-ELM assigns a cluster for each object in a given dataset via the kernel k-means algorithm based on multiple kernels with a set of combination coefficients. It treats the assigned cluster as the pseudo-label for each object. At the second stage, TUMK-ELM learns optimal kernel combination coefficients based on the learned pseudo-label by an analytic solution. This set of coefficients will be further used at the first stage of TUMK-ELM in the next iteration. TUMK-ELM iteratively repeats these two stages until the kernel k-means objective function is converged. Since the time complexity of each stage is small, TUMK-ELM enjoys a high speed of learning from multiple source information.

1.2 Purpose:

ISSN : 2057-5688

A fast unsupervised heterogeneous data learning approach, namely Two-stage Unsupervised Multiple Kernel Extreme Learning Machine (TUMK-ELM, for short). TUMK-ELM iteratively extracts information from multiple sources and learns the heterogeneous data representation with closed-form solutions. It adopts multiple kernels to capture information in heterogeneous data and learns an optimal kernel for heterogeneous data representation. Different from current unsupervised multiple kernel learning methods, it seamlessly integrates a much more efficient kernel combination coefficients optimization method with an effective unsupervised learning objective that simultaneously guarantees a fast learning speed and a high learning quality.

1.3 Scope:

In most real-world data analytics problems, a huge amount of data are collected from multiple sources without label information, which is often with different types, structures, and distributions, namely heterogeneous data.

1.4 Motivation:

One promising way to reveal information from multiple sources is using multiple kernel learning (MKL, for short). MKL first adopts multiple kernels to capture heterogeneous data characteristics from different sources. It then learns optimal combination coefficients for these kernels guided by a specific learning task. In this way, MKL can effectively capture different complex distributions by different kernels, and reveal the relations between these different distributions by the kernel combination coefficients. Despite the advantages of MKL, it requires supervised label information to learn the optimal kernel combination coefficients. However, label information is often not available or very costly in real big data analytics task, which limits the application of MKL.

1.5 Overview:

Unsupervised MK has been studied to tackle the heterogeneous data learning without supervised labels. Similar to MKL, unsupervised MKL also uses multiple kernels to distill information from various sources. To enable the learning without supervised labels, it introduces a kernel-based unsupervised learning objective, e.g. kernel k-means, to learn the optimal kernel combination coefficients. Although unsupervised MKL achieves remarkable performance in unsupervised heterogeneous

ISSN : 2057-5688

data learning, most of the current unsupervised MKL methods are with a slow learning speed. The slow learning speed is mainly caused by the iterative numerical solution, which is adopted by these methods for optimizing the kernel combination coefficients. It does not satisfy the requirements of(1)handling a large amount of data and (2) real-time learning.

2. LITERATURE SURVEY

This work is most related to two learning paradigms. The one is unsupervised deep learning that utilizes deep models to handle large data complexities. The other one is unsupervised multiple view learning that leverages heterogeneous information from multiple views/modes.

A. UNSUPERVISED DEEP LEARNING

Recently, lots of efforts have been done for unsupervised deep learning [1], which aims to reveal complex relations/patterns/knowledge in huge amount of data [2]. Typically, the unsupervised deep learning method combines unsupervised objective and deep neural networks to learn a powerful data representation [3]. For example, the methods in [4] adopt the input reconstruction as the unsupervised objective to learn an insight representation of data. To link the representation more related to analytics tasks, some methods use clustering objective and/or distribution divergence as the learning objective [5], because such objectives may induce a representation with a clearer structure. More recently, many efforts try to learn unsupervised data representation in adversarial approaches [6], which simultaneously take the advantages of both deep generator and deep discriminator. Although such unsupervised deep learning methods can capture highly complex patterns and extremely non-linear relations, they cannot learn heterogeneous data well in an unsupervised fashion. The key reason is that heterogeneous data may have much higher complexity and cause the learning methods converge at a local optimum. Without strong supervised information, the deep network may arbitrarily fit the complex heterogeneous data that leads to meaningless solution.

B. UNSUPERVISED MULTIPLE VIEW LEARNING

Unsupervised multiple view learning aims to learn heterogeneous data without supervised information [7]. Among various unsupervised multiple view learning

ISSN : 2057-5688

methods, unsupervised multiple kernel learning methods attract the most attention because of their ability to represent highly complex data with multimodality. The unsupervised multiple kernel learning is first proposed in [8]. After that, the work in [9] adaptively changes multiple kernel combination coefficients to better capture localized data characteristics. To enhance the robustness of the unsupervised multiple kernel learning, the work in [10] introduces a 2:1-norm to regularize the space of kernel combination coefficients. More recently, [11] proposes local kernel alignment methods to focus on local data relationships. Although the above methods achieve remarkable performance in terms of heterogeneous data representation, all of them fail to apply in big data analytics tasks due to lack of efficiency.

3. EXISTING SYSTEM

More recently, unsupervised MKL has been studied to tackle the heterogeneous data learning without supervised labels. Similar to MKL, unsupervised MKL also uses multiple kernels to distill information from various sources. To enable the learning without supervised labels, it introduces a kernel-based unsupervised learning objective, e.g. kernel k-means, to learn the optimal kernel combination coefficients. Although unsupervised MKL achieves remarkable performance in unsupervised heterogeneous data learning, most of the current unsupervised MKL methods are with a slow learning speed.

3.1 Disadvantages

The slow learning speed is mainly caused by the iterative numerical solution, which is adopted by these methods for optimizing the kernel combination coefficients. It does not satisfy the requirements of (1) handling a large amount of data and (2) real-time learning.

4. PROPOSED SYSTEM

To address the above issues, this work propose a fast unsupervised heterogeneous data learning approach, namely Two-stage Unsupervised Multiple Kernel Extreme Learning Machine (TUMK-ELM, for short). TUMK-ELM iteratively extracts information from multiple sources and learns the heterogeneous data representation with closed-form solutions. It adopts multiple kernels to capture information in heterogeneous data and learns an optimal kernel for heterogeneous data representation.

4.1 Advantages

ISSN : 2057-5688

For extracting knowledge from such big unlabeled heterogeneous data, advanced unsupervised learning techniques are required to (1) have a large model capacity/complexity, (2) have the ability to integrating information from multiple sources and (3) have a high learning speed. The proposed TUMK-ELM has all these requirements

5. Architecture

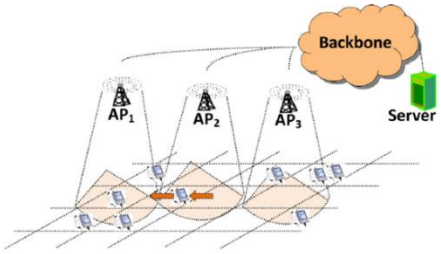


Fig. 1. System architecture: A collection of APs provides service to mobile users according to their locations. Each AP may cover more than one region.

6. IMPLEMENTATION

6.1 Multiple Kernels:

TUMK-ELM first projects heterogeneous data into kernel spaces by multiple kernels. It then adopts an iterative two stages approach to integrate heterogeneous information.

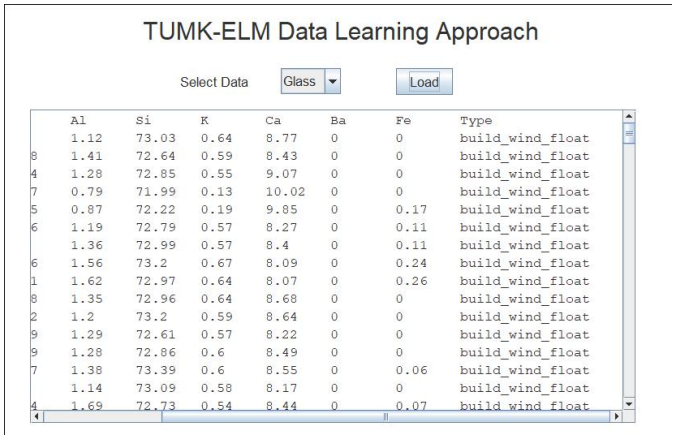
6.2 Stage 1: K-Space:

At the first stage, TUMK-ELM generates a K-Space, in which the data is constructed from multiple kernel spaces and the pseudo-labels are assigned according to the learned optimal kernel.

6.3 Stage 2: Optimal Kernel Combination Coefficients:

At the second stage, TUMK-ELM learns optimal kernel combination coefficients based on the generated K-Space. After convergence, the optimal kernel contains the integrated information from heterogeneous data that suits for the subsequent analytics tasks.

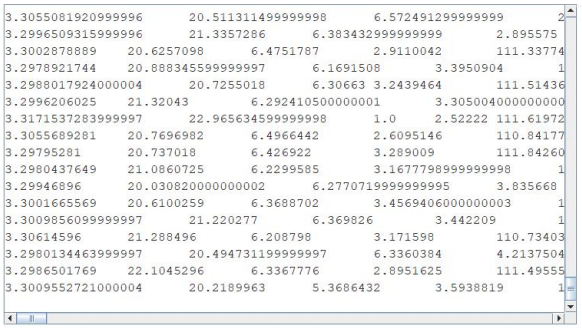
7.OUTPUT RESULTS



	Al	Si	K	Ca	Ba	Fe	Type
	1.12	73.03	0.64	8.77	0	0	build_wind_float
8	1.41	72.64	0.59	8.43	0	0	build_wind_float
4	1.28	72.85	0.55	9.07	0	0	build_wind_float
7	0.79	71.99	0.13	10.02	0	0	build_wind_float
5	0.87	72.22	0.19	9.85	0	0.17	build_wind_float
6	1.19	72.79	0.57	8.27	0	0.11	build_wind_float
	1.36	72.99	0.57	8.4	0	0.11	build_wind_float
6	1.56	73.2	0.67	8.09	0	0.24	build_wind_float
1	1.62	72.97	0.64	8.07	0	0.26	build_wind_float
8	1.35	72.96	0.64	8.68	0	0	build_wind_float
2	1.2	73.2	0.59	8.64	0	0	build_wind_float
9	1.29	72.61	0.57	8.22	0	0	build_wind_float
9	1.28	72.86	0.6	8.49	0	0	build_wind_float
7	1.38	73.39	0.6	8.55	0	0.06	build_wind_float
	1.14	73.09	0.58	8.17	0	0	build_wind_float
4	1.69	72.73	0.54	8.44	0	0.07	build_wind_float

ISSN : 2057-5688

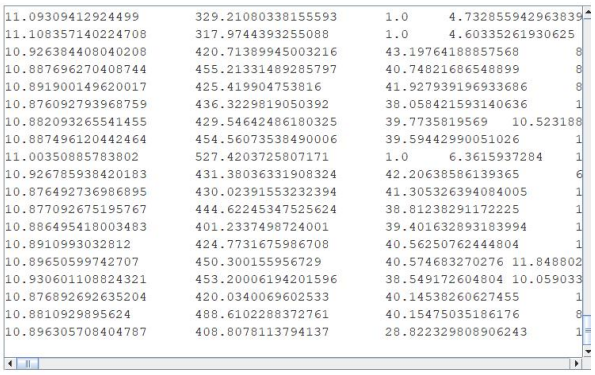
Linear Kernel Matrix



3.3055081920999996	20.511311499999998	6.572491299999999	2	
3.2996509315999996	21.3357286	6.383432999999999	2.895575	
3.3002878889	20.6257098	6.4751787	2.9110042	111.33774
3.2978921744	20.888345599999997	6.1691508	3.3950904	1
3.2988017924000004	20.7255018	6.30663	3.2439464	111.51436
3.2996206025	21.32043	6.292410500000001	3.3050040000000000	
3.3171537283999997	22.965634599999998	1.0	2.52222	111.61972
3.3055689281	20.7696982	6.4966442	2.6095146	110.84177
3.29795281	20.737018	6.426922	3.289009	111.84260
3.2980437649	21.0860725	6.2299585	3.1677798999999998	1
3.29946896	20.030820000000002	6.2770719999999995	3.835668	
3.3001665569	20.6100259	6.3688702	3.4569406000000003	1
3.3009856099999997	21.220277	6.369826	3.442209	1
3.30614596	21.288496	6.208798	3.171598	110.73403
3.2980134463999997	20.494731199999997	6.3360384	4.2137504	
3.2986501769	22.1045296	6.3367776	2.8951625	111.49555
3.3009552721000004	20.2189963	5.3686432	3.5938819	1

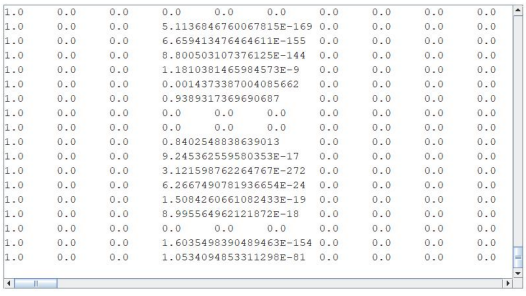
Polynomial Kernel

Polynomial Kernel Matrix



11.09309412924499	329.21080338155593	1.0	4.732855942963839
11.108357140224708	317.9744393255088	1.0	4.60335261930625
10.926384408040208	420.71389945003216	43.19764188857568	8
10.887696270408744	455.21331489285797	40.74821686548899	8
10.891900149620017	425.419904753816	41.927939196933686	8
10.876092793968759	436.3229819050392	38.058421593140636	1
10.882093265541455	429.54642486180325	39.7735819569	10.523188
10.887496120442464	454.56073538490006	39.59442990051026	1
11.00350885783802	527.4203725807171	1.0	6.3615937284 1
10.926785938420183	431.38036331908324	42.20638586139365	6
10.876492736986895	430.02391553232394	41.305326394084005	1
10.877092675195767	444.62245347525624	38.81238291172225	1
10.886495418003483	401.2337498724001	39.401632893183994	1
10.8910993032812	424.7731675986708	40.56250762444804	1
10.89650599742707	450.300155956729	40.574683270276	11.848802
10.930601108824321	453.20006194201596	38.549172604804	10.059033
10.876892692635204	420.0340069602533	40.14538260627455	1
10.8810929895624	488.6102288372761	40.15475035186176	8
10.896305708404787	408.8078113794137	28.822329808906243	1

Gaussian Kernel Matrix



1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	5.113684676006781 5E-169			0.0	0.0	0.0	0.0
1.0	0.0	0.0	6.659413476464611E-155			0.0	0.0	0.0	0.0
1.0	0.0	0.0	8.800503107376125E-144			0.0	0.0	0.0	0.0
1.0	0.0	0.0	1.1810381465984573E-9			0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0014373387004085662			0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.9389317369690687			0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.8402548838639013			0.0	0.0	0.0	0.0
1.0	0.0	0.0	9.245362559580353E-17			0.0	0.0	0.0	0.0
1.0	0.0	0.0	3.121598762264767E-272			0.0	0.0	0.0	0.0
1.0	0.0	0.0	6.266749078193665 4E-24			0.0	0.0	0.0	0.0
1.0	0.0	0.0	1.5084260661082433E-19			0.0	0.0	0.0	0.0
1.0	0.0	0.0	8.995564962121872E-18			0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	1.6035498390489463E-154			0.0	0.0	0.0	0.0
1.0	0.0	0.0	1.0534094853311298E-81			0.0	0.0	0.0	0.0

8 CONCLUSION AND FUTURE ENHANCEMENTS

This work has proposed a Two-Stage Unsupervised multiple kernel Extreme Learning Machines (TUMK-ELM), a more flexible algorithm for fast unsupervised heterogeneous data learning. According to the experiments, the learning speed can achieve as much as 1,000 times faster than RMKKM, 140,000 times faster than LMKKM, and 8,500 times faster than MKC-LKAM. Meanwhile, the clustering accuracy of our proposed TUMK-ELM is comparable with its competitors. Experimental results clearly demonstrate the superiority of TUMK-ELM. In the future, how to adaptive adjust the base kernels to fit the dynamic heterogeneous data distributions will be considered.

9. BIBLIOGRAPHY

[1] Y. LeCun, Y. Bengio, and G. Hinton, ‘‘Deep learning,’’ Nature, vol. 521, pp. 436–444, May 2015.

[2] H. Lee, Y. Largman, P. Pham, and A. Y. Ng, ‘‘Unsupervised feature learning for audio classification using convolutional deep belief networks,’’ in Proc. 22nd Int. Conf. Neural Inf. Process. Syst., 2009, pp. 1096–1104.

[3] Y. Bengio, A. Courville, and P. Vincent, ‘‘Representation learning: A review and new perspectives,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

ISSN : 2057-5688

[4] L. Deng, M. L. Seltzer, D. Yu, A. Acero, A.-R.Mohamed, and G. Hinton, ‘‘Binary coding of speech spectrograms using a deep auto-encoder,’’ in Proc. 11th Annu.Conf. Int. Speech Commun. Assoc., 2010, pp. 1692–1695.

[5] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, ‘‘Deep clustering: Discriminative embeddings for segmentation and separation,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2016, pp. 31–35.

[6] I. Goodfellow et al., ‘‘Generative adversarial nets,’’ in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2672–2680.

[7] B. Long, P. S. Yu, and Z. Zhang, ‘‘A general model for multiple view unsupervised learning,’’ in Proc. SIAM Int. Conf. Data Mining, 2008, pp. 822–833.

[8] S. Yu et al., ‘‘Optimized data fusion for kernel K-means clustering,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 5, pp. 1031–1039, May 2012.

[9] M. Gönen and A. A. Margolin, ‘‘Localized data fusion for kernel K-means clustering with application to cancer biology,’’ in Proc. 27th Int. Conf. Neural Inf. Process. Syst., vol. 1, 2014, pp. 1305–1313.

[10] L. Du et al., ‘‘Robust multiple kernel K-means using ‘2;1-norm,’’ in Proc. Int. Conf. Artif. Intell., 2015, pp. 3476–3482.

ABOUT AUTHORS:

BANDI SARIRAM is currently pursuing MCA in SVKP & Dr K S Raju Arts& Science College(A), affiliated to AdikaviNannaya University, Rajamahendravaram. His research interests include Operation Research, Design and Analysis of Algorithm and Big Data Analytics.

ISSN : 2057-5688

P.SRINIVASA REDDY is working as Associate Professor in SVKP & Dr K S Raju Arts & Science College(A), Penugonda , West Godavari District, A.P. He received Master's Degree in Computer Applications from Andhra University. His research interests include Operational Research, Probability and Statistics, Design and Analysis of Algorithm, Big Data Analytics.