УДК 681.3

E. A. Engel, I. V. Kovalev

INFORMATION PROCESSING USING INTELLIGENT ALGORITHMS BY SOLVING WCCI 2010 TASKS

The article focused on the urgent problem of selection of strategies to deal with ill-structured problems involving the processing of both quantitative and qualitative data, high dimensionality and omissions in the data.

This article provides a detailed analysis of the prediction models for data processing. Experiments confirm the effectiveness of intelligent algorithms, developed by the authors.

Keywords: data processing, ill-structured problem, intelligent algorithms.

Classification problems are found in many application domains, including classification of images or videos, speech recognition, medical diagnosis, marketing, and text categorization.

The category identifiers are referred to as „labels“. Predictive models capable of classifying new instances (correctly predicting their labels) usually require „training“, or parameter adjustment, with large amounts of labeled training data (pairs of examples of instances and associated labels). Unfortunately, few labeled training data may be available due to the cost or burden of manually annotating data. Labeling data is not only expensive, it is tedious. In recent years, Amazon Mechanical Turk and other crowd-sourcing platforms have emerged as a way of rapidly labeling large datasets. However, these are not appropriate for personal or sensitive data. To help us quickly tag our personal pictures, videos, and documents, we need systems that can learn with very few training examples. „Active learning“ helps reducing the burden of labeling by letting the learning machine query only the examples for which the labels are informative.

Following the seminal work in multi-task learning [1], there has been considerable progress in the past decade in developing cross-task transfer using both discriminative and generative approaches in a wide variety of settings [2]. These approaches include multi-layer structured learning machines from the „Deep Learning“ family (Convolutional neural networks, Deep Belief Networks, Deep Boltzmann Machines) [3–6], sparse coding [7–8], and matrix factorization methods, metric or kernel learning methods [9–13]. „Learning to learn“ new concepts [14] is a promising area of research in both machine learning and cognitive science revolving around these ideas. Important progress has also been made in purely unsupervised learning [15–19].

Brief overview of Unsupervised and Transfer Learning. Intelligent beings commonly transfer previously learned knowledge to new domains, making them capable of learning new tasks from very few examples. In contrast, many approaches to machine learning have been focusing on „brute force“ supervised learning from massive amounts of labeled data. While this approach is practical when such data are available, it does not apply when the available training data are mostly unlabeled. Furthermore, even when large amounts of labeled data are available, some categories may be underrepresented. There are many applications for which it would be desirable to learn from very few examples, including just one (one shot learning). The classification accuracy of classifiers trained with very few examples largely depends on the quality of the data representation.

In their review, Pan and Yang [2] give the following definitions: Semi-supervised learning addresses the problem that the labeled data may be too scarce to build a good classifier, by making use of a large amount of unlabeled data and a small amount of labeled data. Transfer learning, in contrast, allows the domains, tasks, and distributions used in training and testing to be different. Transfer learning systems recognize and apply knowledge and skills learned in previous tasks to novel tasks.

Within this framework, there are a variety of settings [2], depending on whether:

– labels are available in the source domain and/or the target domain;

– the tasks are the same or different in the source domain and target domain.

Figure 1, adapted from [2], represents the various situations addressed in the literature.

Unsupervised methods provides an array of possibilities for learning new representations, including:

1) dimensionality reduction or manifold learning;

2) clustering;

3) latent variable or generative models learning.

Principal Component Analysis (PCA), is a method of (linear) projection into a subspace of lower dimension spanned by the eigenvectors of the covariance matrix corresponding to the largest eigenvalues. By construction, the basis vectors in PCA are orthogonal. Other methods for linear dimensionality reduction compute basis vectors in a different way. For instance, Independent Component Analysis (ICA) seeks basis vectors that are independent. Both PCA and ICA can be „kernelized“, to obtain non-linear transformations. Other methods seek transformations into lower dimensional spaces that preserve the local topology (e. g., Kohonen maps, MDS, Isomap, LLE, Laplacian Eigenmaps). Many such methods can be regrouped under the general framework of „regularized principal manifolds“ [15–16] or graphical latent variable models [18]. Among clustering methods [20], *k*-means clustering is the simplest and most widely used.

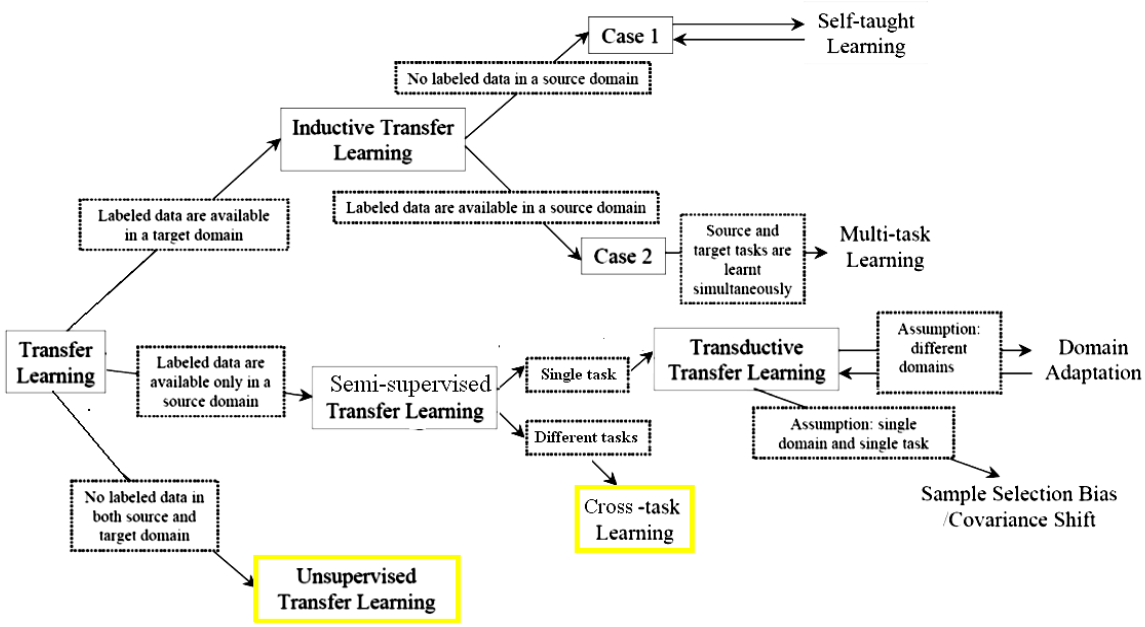

Fig. 1. A taxonomy of transfer learning settings

Starting from k randomly selected cluster centers, it iteratively refines the position of the centers by alternating two steps: (1) forming clusters by assigning examples to their closest cluster center, and (2) re-computing each cluster center by averaging the examples in each cluster. There are many methods related to k-means, which model overlapping clusters, including Gaussian mixtures and fuzzy C-means. One way of exploiting clustering for feature construction is to associate each cluster center with a feature that measures the similarity of the example to that cluster center. While k-means and related methods seek a „shallow“ latent structure in data, hierarchical clustering instead works under the assumption that clusters may be organized in a (deep) hierarchy. The most commonly used hierarchical methods recursively group clusters starting from single examples (bottom up agglomerative methods). Other methods of clustering include graph partitioning and spectral clustering [19].

The successful application of „shallow“ architectures, like kernel methods, has driven away the focus of attention from multi-layer structures (multi-layer neural networks, Deep Belief Networks, Deep Boltzmann Machines, and deep Bayesian latent variable models), which could potentially learn more accurate classifiers for more complex problems, but are more difficult to optimize. However, one of the recent developments in Deep Learning research has been the invention of new algorithms for learning internal representations using unsupervised learning [4–6]. A simple neural network implementation of PCA is the linear „autoencoder“: a three layer neural network of neurons having a linear activation function, in which the output layer tries to reproduce the input layer and the hidden units learn the top principal components by least-square optimization.

Transfer Learning may address in one of two ways:

1) metric, similarity or kernel learning;

2) data representation learning.

There is a wide variety of methods of metric learning or similarity learning (see e. g., [12–13], for a review). We use the notion similarity learning, for algorithms that learn a similarity matrix, which is symmetric but not necessarily.

While kernel learning has been developed relatively recently [10; 11], methods for similarity learning with neural networks have been in use for almost two decades. The idea [9] is to use two replicas of the same neural network, constrained to share parameters. The inputs to the two neural networks are two instances to be compared. The outputs of the networks are combined with a simple parameter-free similarity function such as the cosine of the two output vectors to provide a similarity score. The network is trained to give a large similarity score to examples of the same class and a low score to examples of different classes. Data representation learning is also a landmark of neural networks: in transfer learning, data representations obtained by learning a source task may be re-used in full or in parts to train a system on a target task [5].

Evaluation. *Score: the Area under the Learning Curve.* A learning curve plots the AUC as a function of the number of training examples. We consider two baseline learning curves:

– the ideal learning curve, obtained when perfect predictions are made (AUC = 1). It goes up vertically then follows AUC = 1 horizontally. It has the maximum area „Amax“;

– the „random“ learning curve, obtained by making random predictions (expected value of AUC: 0,5). It follows a straight horizontal line. Its area „Arand“.

To obtain ranking score, we normalize the ALC as follows:

$$globalscore = (ALC - Arand)/(Amax - Arand).$$

We interpolate linearly between points. The score depends on how we scale the *x*-axis. We use a log2 scaling.

Classifier used. We use a linear discriminat classifier to evaluate the quality of the data representations. Denoting by $w = [w_1;w_2; : : : ;w_n]$ the parameter vector of the model, classification is performed using the discriminant function

$$f(x) = w \cdot x. \quad (1)$$

If a threshold is set, patterns having a discriminant function value exceeding the threshold are classified in the positive class. Otherwise they are classified in the negative class. The weights w_i are computed as the difference between the average of feature x_i for the examples of the positive class and the average of feature x_i for the examples of the negative class.

The Area under the ROC Curve (*AUC*). The AUC is the area under the curve plotting *sensitivity* vs. (1 – *specificity*) when the threshold θ is varied (or equivalently the area under the curve plotting *sensitivity* vs. *specificity*).

The results of classification, obtained by thresholding the prediction score, may be represented in a confusion matrix (Table 1), where *tp* (true positive), *fn* (false negative), *tn* (true negative) and *fp* (false positive) represent the number of examples falling into each possible outcome. We define the sensitivity (also called true positive rate or hit rate) and the specificity (true negative rate) as:

$$Sensitivity = tp/pos;$$
$$Specificity = tn/neg$$

where $pos = tp + fn$ is the total number of positive examples and $neg = tn + fp$ the total number of negative examples.

Table 1

Confusion matrix

Truth		Prediction	
		Class + 1	Class – 1
	Class + 1	*tp*	*fn*
	Class – 1	*fp*	*tn*

We then estimate the standard deviation of the *BAC* as

$$\sigma = \frac{1}{2}\sqrt{\frac{p_+(1-p_+)}{pos} + \frac{p_-(1-p_-)}{neg}}, \quad (2)$$

where *pos* is the number of examples of the positive class, *neg* is the number of examples of the negative class, and p_+ and p_- are the probabilities of error on examples of the positive and negative class respectively, approximated by their empirical estimates, the *sensitivity* and the *specificity* (see Figure 2) [21].

Experimental Results. The modified neural network [22] solves practical tasks in various subject fields. To investigate the generality of the modified artificial neural network we solved in different domains: handwriting recognition, marketing and chemoinformatics. This section reports the results of numerical experiments which indicate, that algorithms proposed by author [22–24] has appropriate generalization accuracy.

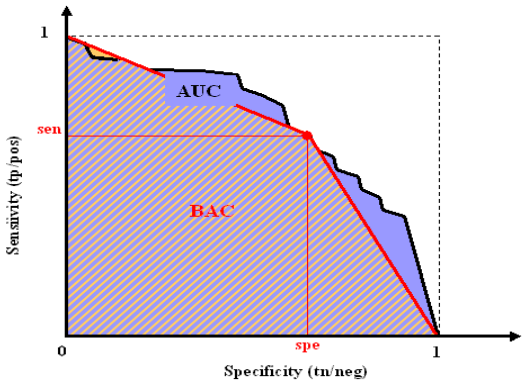

Fig. 2. The Area under the ROC Curve (AUC)

Handwriting recognition: IBN_SINA. Historical archive collections are difficult to process by traditional Optical Character Recognition (OCR) methods, due to their historical character types or due to the fact that the material is handwritten and uses scripts that are no longer in use. There are thousands of different scripts in use worldwide and large volumes of scanned documents waiting to be indexed to facilitate retrieval. Transfer learning methods could accelerate the application of handwriting recognizers by reducing the need for using human experts to label data.

IBN_SINA is a handwriting recognition dataset. The task of IBN_SINA is to spot arabic words in an ancient manuscript to facilitate indexing. The data were formatted in a feature representation (92 variables).

The task of IBN_SINA solved by using the modified neural network with detector-descriptor scheme as preprocessing step [23]. First we obtain database using detector-descriptor scheme on the IBN_SINA database. For detector and descriptor we using the testing software provided by Mikolajczyk. The matching is carried out as follows. There were 1000 3-layer modified artificial neural networks (33 hidden units: 30 and 3 units at 1st and 2nd hidden layer respectively) trained on the preprocessing IBN_SINA database. Our ranking is 8, done according to the Score and compare with the WCCI results on IBN_SINA (Table 2).

The preprocessing step with detector-descriptor scheme increased Score of modified artificial neural network from 0.913152 to 0.977364.

ORANGE. ORANGE is a marketing dataset. Customer Relationship Management (CRM) is a key element of modern marketing strategies. This dataset was extracted from a large marketing database from the French Telecom company Orange.

The goal is to predict the propensity of customers to switch provider (churn), buy new products or services (appetency), or buy upgrades or add-ons proposed to them to make the sale more profitable (up-selling).

Table 2

Ranking on WCCI 2010 tasks

Ranking	IBN_SINA Score (AUC)	ORANGE Score (AUC)	HIVA Score (AUC)
1	0,990 445	0,810 102	0,947 128
2	0,988 359	0,721 316	0,884 041
3	0,983 816	0,813 333	0,812 109
4	0,978 695	0,787 346	0,793 947
5	0,977 876	0,788 271	0,784 561
6	0,97 781	0,78 821	0,770 527
7	0,977 415	0,787 634	0,769 719
8	0,977 364	0,787 244	0,711 866
9	0,977 071	0,813 333	0,681 118

The difficulties include heterogeneous noisy data (numerical and categorical variables), and unbalanced class distributions. We provide below the list of categorical variables:

3 10 16 25 27 32 33 47 49 59 65 73 75 76 79 81 88 96 98 100 105 112 113 121 128 132 138 140 141 148 152 153 154 167 173 181 187 194 209 216.

The task of ORANGE solved by using the fuzzy neural network [24]. First we obtain database using numerical representation of categorical variables on the ORANGE database. There were 1000 4-layer fuzzy neural networks (53 hidden units: 40, 10 and 3 units at 1st, 2nd and 3nd hidden layer respectively) trained on the preprocessing ORANGE database. Our ranking is 7, done according to the Score and compare with the WCCI results on ORANGE (Table 2).

Experimental results show that fuzzy neural network performs quite well compare to other algorithms.

HIVA. HIVA is a chemoinformatics dataset. The task of HIVA is to predict which compounds are active against the AIDS HIV infection. The original data has 3 classes (active, moderately active, and inactive). We brought it back to a two-class classification problem (active vs. inactive). We represented the data as 1617 sparse binary input variables. The variables represent properties of the molecule inferred from its molecular structure. The problem is therefore to relate structure to activity (a QSAR = quantitative structure-activity relationship problem) to screen new compounds before actually testing them (a HTS = high-throughput screening problem).

The original data were made available by The National Cancer Institute (USA). The 3d molecular structure was obtained by the CORINA software and the features were derived with the ChemTK software.

The HIVA dataset was used previously in the Performance Prediction challenge, the Model Selection game, and the Agnostic Learning vs. Prior Knowledge (ALvsPK) challenge. A variant of the HIVA dataset called SIDO was used in the Causation and Prediction challenge and the Pot-Luck challenge.

The task of HIVA solved by using the modified neural network [22]. There were 1000 3-layer modified artificial neural networks (57 hidden units: 50 and 7 units at 1st and 2nd hidden layer respectively) trained on the HIVA database. Our ranking is 5, done according to the Score and compare with the WCCI results on HIVA (Table 2).

Experimental results show that fuzzy neural network performs very well compare to other algorithms.

We have investigated algorithms proposed by author [22–24] which fulfills the optimal complex and cross-validated model. Our analysis was based on object recognition and classification tasks. The algorithms developed by author applied to solved tasks in different domains: handwriting recognition, marketing and chemoinformatics. Experimental results show that:

– the modified artificial neural network effectively solves practical tasks of various subject fields and performs very well compare to the popular learning algorithms and advisable to gain extra prediction accuracy;

– the fuzzy neural network have done well to predict performance of tasks with categorical variables;

– at handwriting recognition tasks [23] the detector-descriptor scheme as preprocessing step significantly improve performance of modified artificial neural network.

References

1. Caruana R. Multitask learning // Machine Learning. 1997. Vol. 28. № 1. P. 41–75.

2. Pan S. J., Yang Q. A survey on transfer learning // IEEE Trans. on Knoweledge and Data Engineering. 2010. Vol. 22. № 10. P. 1345–1359.

3. Collobert R., Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning // Intern. Conf. on Machine Learning (ICML). 2008. P. 160–167.

4. Bengio Y. Learning deep architectures for AI // Foundations and Trends in Machine Learning. 2009. Vol. 2. № 1. P. 1–127.

5. Gutstein S. M. Transfer learning techniques for deep neural nets : Ph. D. dissertation. The University of Texas at El Paso, 2010.

6. Why does unsupervised pre-training help deep learning? / D. Erhan, Y. Bengio, A. Courville et al. // JMLR. 2010. Vol. 11. P. 625–660.

7. Efficient sparse coding algorithms / H. Lee, A. Battle, R. Raina, A. Y. Ng // Advances in Neural Information Processing Systems. 2007. Vol. 19. P. 801–808.

8. Self-taught learning: Transfer learning from unlabeled data / R. Raina, A. Battle, H. Lee et al. // Proc. of the Twenty-fourth Intern. Conf. on Machine Learning, 2007. P. 759–766.

9. Signature verification using a “siamese” time delay neural network / J. Bromley, I Guyon., Y. LeCun et al. // NIPS. 1993. P. 737–744.

10. Learning the kernel matrix with semi-definite programming // G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui // J. of Machine Learning Research. 2004. Vol. 5. P. 27–72.

11. Weinberger K. Q., Saul L. K. Distance metric learning for large margin nearest neighbor classification // J. Machine Learning Research. 2009. Vol. 10 P. 207–244.

12. Yang L., Jin R. Distance metric learning: A comprehensive survey [Electronic resource] : Techn. Rep. Michigan State University. 2006. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.91.4732 (data of visit: 30.07.2011).

13. Yang L. An overview of distance metric learning [Electronic resource] : Techn. Rep. Carnegie Mellon University. 2007. URL: http://www.cs.cmu.edu/~liuy/dist_overview.pdf (data of visit: 30.07.2011).

14. Learning to Learn / S. Thrun, L.Y. Pratt (ed.). Boston, MA : Kluwer Academic Publishers, 1998.

15. Regularized principal manifolds / A. J. Smola, S. Mika, B. Schlkopf, R. C. Williamson // JMLR. 2001. Vol. 1. P. 179–209.

16. Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering / Y. Bengio, J.-F. Paiement, P. Vincent et al. // NIPS. 2003. P. 177–184.

17. Globerson A., Tishby N. Sufficient dimensionality reduction // J. Machine Learning Research. 2003. Vol. 3. P. 1307–1331.

18. Ghahramani Z. Unsupervised Learning // Advanced Lectures in Machine Learning. Lecture Notes in Computer Sci. Berlin : Springer-Verlag, 2004. Vol. 3176. P. 72–112.

19. Luxburg U. A tutorial on spectral clustering // Statistics and Computing. 2007. Vol. 17. P. 395–416.

20. Jain A. K., Murty M. N., Flynn P. J. Data clustering : A review ACM Computing Surveys. 1999. P. 264–323.

21. Performance prediction challenge / I. Guyon, A. Saffari, G. Dror, J. Buhmann // IEEE/INNS conf. IJCNN 2006. Vancouver, Canada, July 16–21. 2006. P. 1649–1656.

22. Engel E. A. Modified artificial neural network for information processing with the selection of essential connections : Ph. D. thesis. Krasnoyarsk, 2004.

23. Engel E. A. Graphic information processing using intelligent algorithms // Vestnik. Sci. J. of Siberian State Aerospace Univ. № 4(25). 2009. P. 85–90.

24. Engel E. A. The hierarchical model of decision-making based on fuzzy neural networks for information processing, Vestnik. Sci. J. of Siberian State Aerospace Univ. № 1 (33). 2011. P. 83–86.

Е. А. Энгель, И. В. Ковалев

ИСПОЛЬЗОВАНИЕ ИНТЕЛЛЕКТУАЛЬНЫХ МЕТОДОВ ДЛЯ ОБРАБОТКИ ИНФОРМАЦИИ НА ПРИМЕРЕ РЕШЕНИЯ ЗАДАЧ WCCI 2010

Рассмотрена актуальная проблема выбора стратегии решения слабоформализованных задач, предполагающих обработку как количественных, так и качественных данных, высокую размерность и пропуски в данных.

Представлен детальный анализ моделей прогноза для обработки данных. Эксперименты подтверждают эффективность интеллектуальных алгоритмов, разработанных авторами.

Ключевые слова: обработка данных, слабоформализованные задачи, интеллектуальные алгоритмы.

УДК 517.95

M. B. Frost

DETERMINING THE SOURCE OF TRANSVERSE OSCILLATIONS OF AN ELASTIC ROD

Solvability of an inverse problem for the equations of transverse oscillations of an elastic rod (determining the source of oscillations on the basis of the rod deflection at the final time) is proved.

Keywords: elastic rod, inverse problem, source of oscillations, solvability.

Many issues of engineering, geophysics, and medicine involve problems whose sought quantities are elements of initial-boundary problems for differential equations. These unknown elements are determined on the basis of additional information. Such problems are called inverse problems for differential equations. The advanced theory of inverse problems and numerous publications can be found in [1].

Let us consider an inverse problem of determining the source of transverse oscillations of an elastic rod. The initial-boundary problem of transverse oscillations of a simply supported rod with a constant cross section and a