World Applied Sciences Journal 31 (Applied Research in Science, Engineering and Management): 69-75, 2014
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2014.31.arsem.516

Application of Classification Restricted Boltzmann Machine to Medical Domains

Jakub M. Tomczak

Institute of Computer Science, Wroclaw University of Technology,
wyb. Wyspianskiego 27, 50-370, Wroclaw, Poland

Abstract: Recent developments have demonstrated deep models to be very powerful generative models which are able to extract features automatically and obtain high predictive performance. Typically, a building block of a deep architecture is Restricted Boltzmann Machine (RBM). In this work, we focus on a variant of RBM adopted to the classification setting, which is known as Classification Restricted Boltzmann Machine. We claim that this model should be used as a stand-alone non-linear classifier which could be extremely useful in medical domains. Additionally, we show how to obtain sparse representation in RBM by adding a regularization term to the learning objective which enforces sparse solution. The considered classifier is then applied to five different medical domains.

Key words: Restricted boltzmann machine . classification . sparse . medical domain . diabetes . oncology

INTRODUCTION

Deep learning paradigm becomes a crucial part of modern machine learning methods because it allows to extract features automatically and obtain high predictive performance [1]. A building block of a deep architecture could be a probabilistic model called Restricted Boltzmann Machine (RBM), used to represent one layer of the deep structure. Restricted Boltzmann Machines are interesting because they are capable of learning complex featuers and they form a bipartie graph which makes inference easy in them. Moreover, it has been proven that in the sense of Kullback-Leibler divergence RBM is able to arbitrarily well approximate any distribution over binary inputs for properly chosen number of hidden units [11]. Basing on this result it has been proposed to add an additional layer to RBM representing an output variable. In other words, it has been advocated to use RBM as a stand-alone non-linear classifier [8].

Learning flexible classifiers, which are capable of extracting features in an automatic manner and capturing non-linear dependencies, is especially important in medical domain [7]. It has been shown that such classifiers can be very helpful in medical decision support systems, e.g., Boosted SVM for lung cancer patients [21], Graph-based Rule Inducer for diabetes treatment [19], Bagging of decision trees for breast cancer recurrence [17]. Recently, RBM for classification was applied to breast cancer recurrence and in comparison to other methods it obtained very promising results [18]. In this paper, we aim at verifying the applicability of RBM as a predictive model (or diagnostic tool) in five various medical domains.

The typical learning algorithm for RBM is \textit {Stochastic Gradient Descent} (SGD) technique which scales well for large data and ensures convergence at rate equal to inverse of the desired accuracy (under mild conditions) [3]. However, there could be two problems with learning RBM using SGD. First, the model could be overcomplete [13], i.e., there are too many hidden units. Second, learnt features could be not enough discriminative, i.e., they are suitable for reconstruction but insufficient for prediction. A possible solution to both of these issues is application of sparse learning.

Typical sparse learning is based on adding a regularization term to the learning objective (typically-negative log-likelihood) which penalizes dense solutions. An example of such regularization is L1 norm of model's parameters. It has been shown that this approach allows to learn with overcomplete representations [13] but also leads to worst predictive performance [14]. Another approach is sparse Bayesian learning which aims at introducing latent variable models which enforces sparse solutions [14]. In deep learning there are several methods which introduce sparse solutions. Most of them are based on using regularization term which encourages activation of hidden units to be at low rate. One approach applies cross-entropy between desired activation level and

Corresponding Author: Jakub M. Tomczak, Institute of Computer Science, Wroclaw University of Technology, wyb. Wyspianskiego 27, 50-370, Wroclaw, Poland

estimated probability of activation [15]. Very similar solution proposes to use L2 norm instead of the corss-entropy [10]. Other one applies Bhattacharyya distance in order to differentiate probability of activation among hidden units for given input [12]. In this paper, we give a different view on how to use probability distance measures and show that for some probabilistic similarity measures (e.g. Kullback-Leibler divergence, Bhattacharyya distance, Mahalanobis distance) one obtains the same regularization term.

The paper is organized as follows. In Section II the model of RBM for classification is outlined and its discriminative learning (Section II-B) and sparse learning (Section II-C) are described. In Section III the empirical study is carried out. We compare the RBM for classification with discriminative and sparse learning with well-known classifiers in five medical domains. At the end of the paper (Section IV) conclusions are drawn and directions of future research are indicated.

CLASSIFICATION RESTRICTED BOLTZMANN MACHINE

Classification Restricted Boltzmann Machine (ClassRBM) [8, 9] is a three-layer undirected graphical model where the first layer consists of visible input variables $x \in \{0,1\}^D$, the second layer consists of hidden variables (units) $h \in \{0,1\}^M$ and the third layer represents observable output variable $y \in \{1,2,\ldots,K\}$. We use the 1-to-K coding scheme which results in representing output as a binary vector of length K denoted by y, such that if the output (or class) is k, then all elements are zero except element y_k which takes the value 1. We allow only the inter-layer connections, i.e., there are no connections within layers.

A RBM with M hidden units is a parametric model of the joint distribution of visible and hidden variables, that takes the following form:

$$p(\mathbf{x},\mathbf{y},\mathbf{h}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})}\exp\{-E(\mathbf{x},\mathbf{y},\mathbf{h}|\boldsymbol{\theta})\} \quad (1)$$

with parameters $\theta = \{b, c, d, W^1, W^2\}$ and where:

$$E(\mathbf{x},\mathbf{y},\mathbf{h}|\boldsymbol{\theta}) = -\mathbf{b}^T\mathbf{x} - \mathbf{c}^T\mathbf{h} - \mathbf{d}^T\mathbf{y} - \mathbf{x}^T\mathbf{W}^1\mathbf{h} - \mathbf{h}^T\mathbf{W}^2\mathbf{y} \quad (2)$$

is an energy function and

$$Z(\boldsymbol{\theta}) = \sum_{\mathbf{x},\mathbf{y},\mathbf{h}} \exp\{-E(\mathbf{x},\mathbf{y},\mathbf{h}|\boldsymbol{\theta})\} \quad (3)$$

It can be shown that the following expressions hold true for ClassRBM [8, 9] (Further in the paper, sometimes we omit explicit conditioning on parameters θ)

$$p(\mathbf{x}|\mathbf{h}) = \prod_j p(x_j|\mathbf{h}) \quad (4)$$

$$p(x_j = 1|\mathbf{h}) = \mathrm{sigm}(b_j + \mathbf{W}_j^1\mathbf{h}) \quad (5)$$

$$p(y_k = 1|\mathbf{h}) = \frac{\exp(d_k + (\mathbf{W}_{\cdot k}^2)^T\mathbf{h})}{\sum_l \exp(d_l + (\mathbf{W}_{\cdot l}^2)^T\mathbf{h})} \quad (6)$$

$$p(\mathbf{h}|y_k = 1,\mathbf{x}) = \prod_j p(h_j|y_k = 1,\mathbf{x}) \quad (7)$$

$$p(h_j = 1|y_k = 1,\mathbf{x}) = \mathrm{sigm}(c_j + (\mathbf{W}_{\cdot j}^1)^T\mathbf{x} + W_{jk}^2) \quad (8)$$

where $\mathrm{sigm}(\cdot)$ is the logistic sigmoid function, W_i^ℓ . is the i^{th} row of weights matrix W^ℓ, $W_{\cdot j}^\ell$, is j^{th} column of weights matrix W^ℓ, W_{ij}^ℓ is the element of weight matrix W^ℓ.

Prediction: For given parameters θ it is possible to exactly compute the distribution $p(y|x,\theta)$ which can be further used to choose the most probable class label. This conditional distribution takes the following form [8, 9]:

$$p(y=k|\mathbf{x},\boldsymbol{\theta}) = \frac{\exp(d_k)\prod_j(1+\exp(c_j+(\mathbf{W}_{\cdot j}^1)^T\mathbf{x}+W_{jk}^2))}{\sum_l \exp(d_l)\prod_j(1+\exp(c_j+(\mathbf{W}_{\cdot j}^1)^T\mathbf{x}+W_{jl}^2))} \quad (9)$$

Pre-computing the terms $a_j = c_j + (\mathbf{W}_{\cdot j}^1)^T\mathbf{x}$ allows to reduce the time needed for computing the conditional distribution to O(MD+MK) [8, 9].

Learning: We assume given N data $D = \{x_n, y_n\}$, where n^{th} example consists of an observed inputs x_n and a target class y_n. Typically, learning of a probabilistic model is based on the likelihood function [2]. However, in order to train ClassRBM we may consider two approaches. The first one, called \textit{generative approach}, aims at maximizing the likelihood function for joint distribution $p(y|x,\theta)$. The second one, which we refer to as discriminative approach, considers the likelihood function for conditional distribution $p(y|x,\theta)$. The problem with generative approach is that it is impossible to calculate exact gradient of the likelihood function for joint distribution (only approximation can be applied, e.g., Contrastive Divergence [6]). On the other hand, the latter approach allows to compute exact gradient [9]. Additionally, we are interested in obtaining high predictive accuracy, thus, it is more advantageous to learn ClassRBM in a discriminative

manner. Therefore, to train ClassRBM we consider minimization of the negative log-likelihood in the following form:

$$L(\mathcal{D}|\theta) = -\sum_{n=1}^{N} \log p(\mathbf{y}_n|\mathbf{x}_n, \theta) \tag{10}$$

As stated before, since the distribution $p(y|x,\theta)$ can be calculated exactly, the gradient of (10) can be computed exactly too which yields (Function $1_{a=b}$ i san indicator function which returs 1 if a and b are equal and 0-otherwise):

$$\frac{\partial \log p(\mathbf{y}_n|\mathbf{x}_n)}{\partial a_j} = 1_{y=y_n} - p(y|\mathbf{x}_n), \tag{11}$$

$$\frac{\partial \log p(\mathbf{y}_n|\mathbf{x}_n)}{\partial \theta} = \sum_j \text{sigm}(g_{kj}(\mathbf{x}_n)) \frac{\partial g_{kj}(\mathbf{x}_n)}{\partial \theta} - \sum_{y,j} \text{sigm}(g_{kj}(\mathbf{x}_n)) p(y = 1|\mathbf{x}_n) \frac{\partial g_{kj}(\mathbf{x}_n)}{\partial \theta} \tag{12}$$

Sparse learning: Sparse representations have been shown to be beneficial in practical applications. From the information-theoretic point of view, sparse representations obtain better generalization performance in comparison to non-sparse ones because the training examples should encoded with as small number of bits as possible. On the other hand, learning sparse features helps to improve discriminative capabilities of features, i.e., sparse features become more class-specific.

There are different approaches to introduce sparse representations (see Section 1 for a short review). In this paper, we follow the approach that adds a regularization term to the objective function which enforces sparse solution. Our new objective takes the following form:

$$L_r(\mathcal{D}|\theta) = L(\mathcal{D}|\theta) + \lambda\Omega(\theta) \tag{13}$$

where $\lambda>0$ is a regularization coefficient and $\Omega(\theta)$ is regularization term.

The most popular regularization term is L2 norm of model's parameters which is known as weight decay, i.e., $\Omega(\theta) = \|\theta\|_2^2$. However, we will try to sparsify expected activations of hidden units by forcing them to be kept at fixed small level μ. In order to obtain the desired effect the regularization term could be thought as a measure that compares the difference between two probability distributions. The higher is the difference between expected hidden units activations and μ, the stronger is the regularization. Hence, Ω can be, for example, Kullback-Leibler divergence (KLdiv), Bhattacharyya distance (Bdist), Mahalanobis distance (Mdist)(These three mentioned measures can be calculated analytically for normally distributed random variables, which is not true for other measures, e.g., Kolmogorov distance [20]).

Let us assume that the hidden unit activity is normally distributed with mean equal $p(h_j|x,y)$ and unit variance, i.e., $N(p(h_j|x,y),1)$ and the desired activation level is $N(\mu,1)$ (For KLdiv, Bdist and Mdist the value of variance turns to be a scaling factor which can be further included in the regularization coefficient. Therefore, for simplicity, we choose unit variance). Then, it turns out that application of KLdiv, Bdist or Mdist to our problem yields the following regularization term:

$$\Omega(\theta) \propto \left(\mu - p(h_j|\mathbf{x}_n, \mathbf{y}_n)\right)^2 \tag{14}$$

It is interesting that we have obtained exactly the same regularization term as in [10], i.e., squared difference between μ and $p(h_j|x_n, y_n)$. However, our derivation of the regularization term has formal justification while the result in [10] follows from considerations in neuroscience and is given rather ad hoc as a L2 norm of a difference between expected hidden activations and μ. Moreover, our proposition can be further developed by weakening the assumption about unit variance. However, we leave these investigations for further research.

The objective function for learning is a sum of the negative log-likelihood and the regularization term. Therefore, we can apply stochastic gradient descent algorithm to the new objective (13). For the negative log-likelihood the gradient is given in (11) and (12) and for the regularization term given in (\ref{eq:regularizationTerm}) we get (for simplicity we denote $p(h_j|x_n, y_n)$ by p_{jn}):

$$\frac{\partial\Omega(\theta)}{\partial\theta} \propto (\mu - p_{jn})p_{jn}(1 - p_{jn})\frac{\partial g_{kj}(\mathbf{x}_n)}{\partial\theta} \tag{15}$$

Further, we will refer sparse learning of ClassRBM to as sparseClassRBM.

EXPERIMENT

Setup: During learning ClassRBM and sparseClassRBM for different datasets we kept all parameters fixed. The learning rate was set to 0.001, the number of hidden units was 100. Additionally, we used Nesterov's Accelerated Gradient technique (with parameter's value 0.5) which is special kind of momentum term [16] and a mini-batch of size 100.

ClassRBM and sparseClassRBM were compared with the following classifiers:

- CART,
- AdaBoost,
- LogitBoost,
- Tree Bagging,
- Random Forest,
- SVM,
- NeuralNetwork (three hidden layers).

Datasets: We evaluate ClassRBM and sparseClassRBM on the following medical datasets:

- Heart (joined cleveland and hungarian datasets) [5]: Diagnostic problem of heart disease (less than 50% diameter narrowing in many major vessel or not),
- Diabetes [5]: classification problem of the patient as tested positive for diabetes or not,
- Indian liver [5]: classification problem of the patient as healthy or with a liver issue,
- Sick [5]: diagnostic problem of thyroid disease,
- Oncology [17]: prediction problem of the patient whether there will be a recurrence of breast cancer or not.

The datasets are summarized in Table 1. The number of features and the number of examples for each dataset are given. Additionally, we provide the imbalance ratio defined as the number of majority class examples that are divided by the number of minority class examples.

RESULTS AND DISCUSSION

The results of the experiment are presented in Fig. 1-5 as boxplots. In Table 2 best three methods for each dataset are presented.

It can be noticed that ClassRBM together with sparseClassRBM were the most stable methods which; they four times out of five among best three methods (in the 5th case they were on 4th place). Next three best model were: AdaBoost (three times) and Random Forest and LogitBoost (two times). Therefore, we claim that ClassRBM and its sparse version are strong and robust classifiers. However, they are prone to highly imbalanced data (see poor performance on sick dataset, Fig. 4).

Table 1: The number of features, the number of examples and the imbalance ratio for the medical datasets used in the experiment

Dataset	Number of inputs	Number of examples	Imbalance ration
Heart	46	597	1.45
Diabetes	20	768	1.87
Indian liver	48	583	2.49
Sick	57	3772	15.33
Oncology	55	949	1.60

Table 2: Ranking of classifiers according to Kappa for considered five datasets

Dataset	1st place	2nd place	3rd place
Heart	Sparse class RBM	Logit boost	Class RBM
Diabetes	Class RBM	Sparse class RBM	Ada Boost
Indian liver	Random forest	Sparse RBM	Ada Boost
Sick	Cart	Random forest	Class RBM
Oncology	Logit boost	Ada Boost	Class RBM and Sparse class RBM

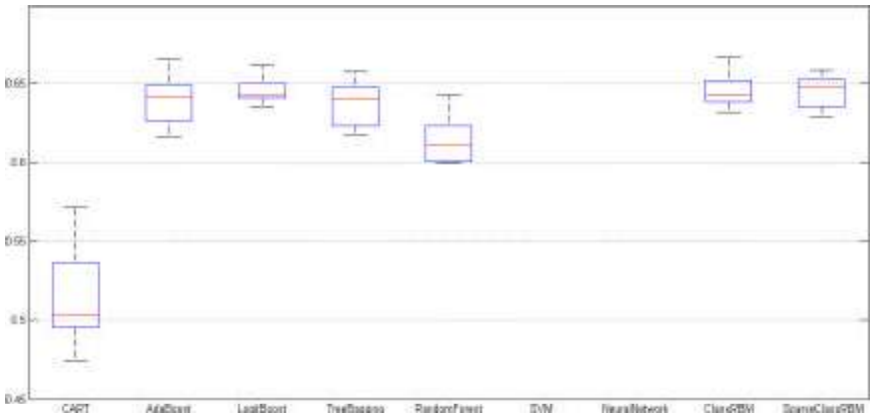

Fig. 1: Boxplot of Kappa values for heart dataset

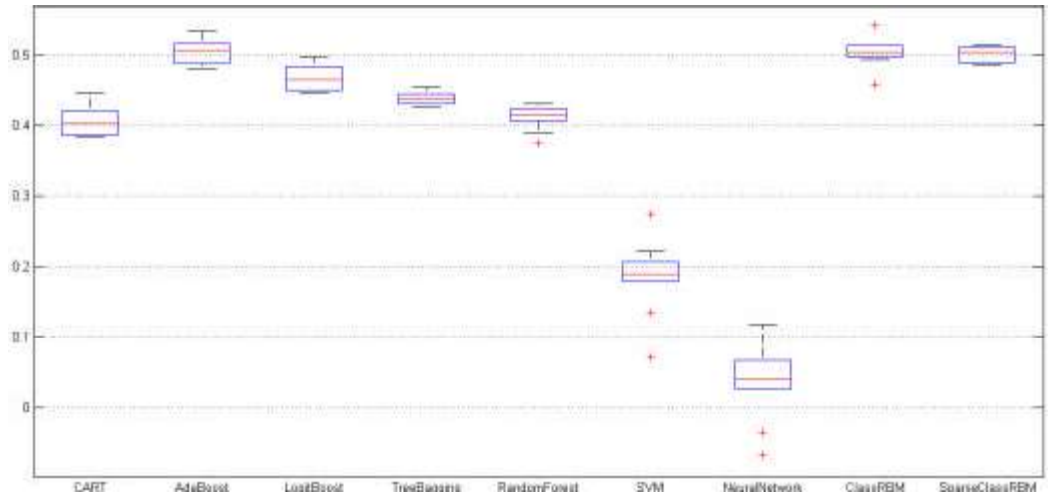

Fig. 2: Boxplot of Kappa values for diabetes dataset

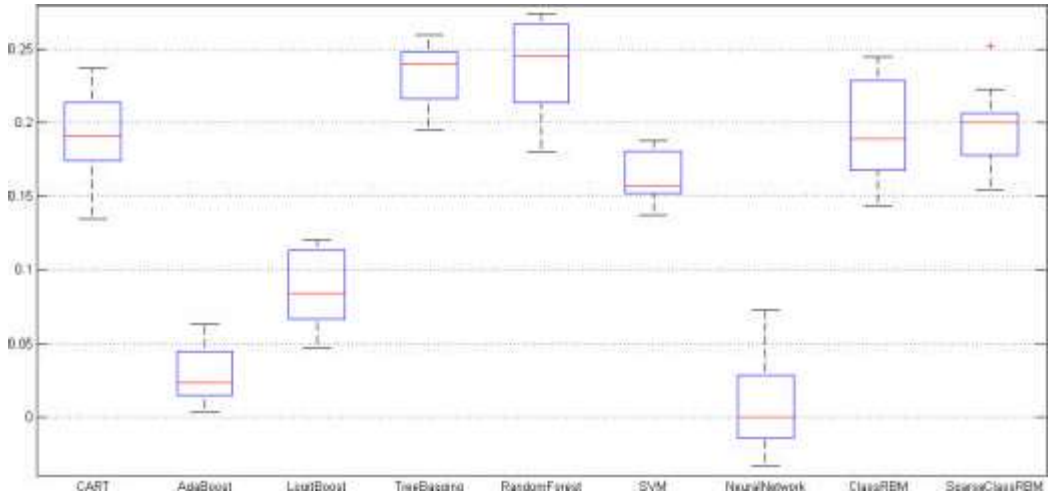

Fig. 3: Boxplot of Kappa values for indian dataset

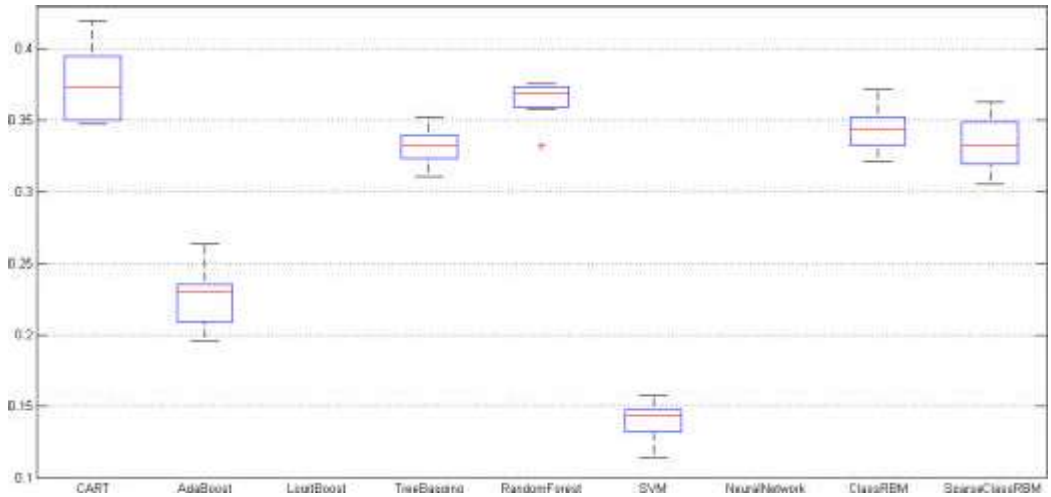

Fig. 4: Boxplot of Kappa values for sick dataset

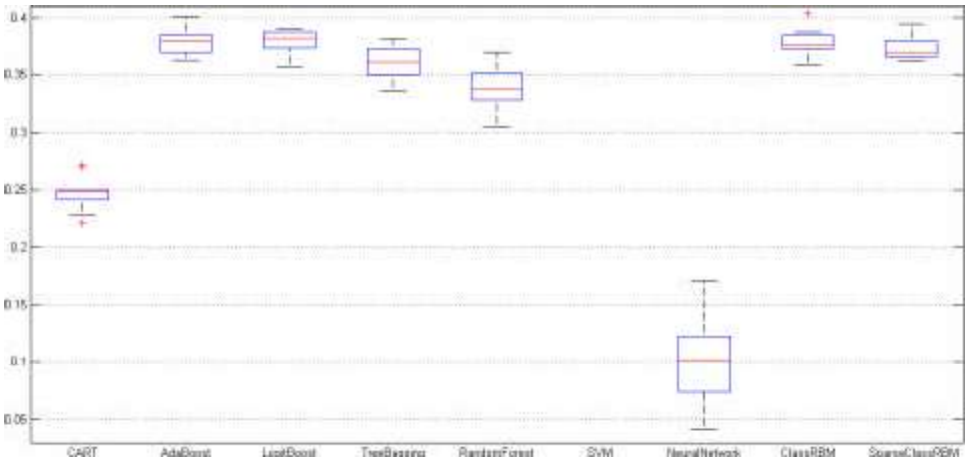

Fig. 5: Boxplot of Kappa values for onko dataset

Comparing ClassRBM to its sparse version basing on the considered datasets is rather inconclusive. It is difficult to tentatively state whether application of sparse regularization term gives better results. However, on heart, diabetes and indian sparsification performes slightly better than ClassRBM, comparably on onko but worst on sick. Nonetheless, we believe that for more hidden units the results would be more conclusive and the regularization term would prevent from overfitting and maybe even improve predictive performance. We leave this aspect for further research.

CONCLUSION

In this paper, we have outlined a deep model called Classification Restricted Boltzmann Machine in application to five medical domains. We follow the way of reasoning given in [8] which says that ClassRBM can be used as stand-alone non-linear classifier. Moreover, we claim that this model is very stable and should be used as a state-of-the-art classifier in any domain and especially in medical domain which demands stable solutions. In the experiments, for five different medical problems, we have shown that both discriminative and sparse learning of ClassRBM give very promising results and outperforms well-known strong classifiers like AdaBoost, LogitBoost, TreeBagging and Random Forest.

In our study two important issues have arisen which have indicated possible future research. First, ClassRBM fails when data are highly imbalanced and thus there is a need to propose some remedy for that issue. Second, in our proposition of sparse learning we have assumed unit variance. However, such approach is very simplistic and weakening this assumption could give very interesting solutions.

ACKNOWLEDGMENT

The research conducted by Jakub M. Tomczak has been partially co-financed by the Ministry of Science and Higher Education, Republic of Poland (grant No. B30098I32).

REFERENCES

1. Bengio, Y., 2009. Learning deep architectures for AI. Foundations and trends in Machine Learning, 2 (1): 1-127.
2. Bishop, C.M., 2006. Pattern recognition and machine learning. New York: Springer.
3. Bottou, L., 2012. Stochastic Gradient Descent Tricks. In: Neural Networks: Tricks of the Trade. Springer Berlin Heidelberg, pp: 421-436.
4. Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 20 (1): 37-46.
5. Frank, A. and A. Asuncion, 2010. UCI machine learning repository.
6. Hinton, G.E., 2002. Training products of experts by minimizing contrastive divergence. Neural Computation, 14 (8): 1771-1800.
7. Kononenko, I., 2001. Machine learning for medical diagnosis: History, state of the art and perspective. Artificial Intelligence in Medicine, 23 (1): 89-109.
8. Larochelle, H. and Y. Bengio, 2008. Classification using discriminative restricted Boltzmann machines. ICML.
9. Larochelle, H., M. Mandel, R. Pascanu and Y. Bengio, 2012. Learning algorithms for the classification restricted Boltzmann machine. The Journal of Machine Learning Research, 13: 643-669.

10. Lee, H., C. Ekanadham and A. Ng, 2007. Sparse deep belief net model for visual area V2. In: Advances in Neural Information Processing Systems, pp: 873-880.
11. Le Roux, N. and Y. Bengio, 2008. Representational power of restricted Boltzmann machines and deep belief networks. Neural Computation, 20 (6): 1631-1649.
12. Le Roux, N., H. Larochelle and Y. Bengio, 2008. Discriminative Training of RBMs using Bhattacharyya Distance. Learning Workshop, Cliff Lodge, Snowbird, Utah.
13. Lewicki, M.S. and T.J. Sejnowski, 1998. Learning nonlinear overcomplete representations for efficient coding. Advances in Neural Information Processing Systems, pp: 556-562.
14. Mohamed, S., Ghahramani, Z., Heller, K.A. Bayesian and L1, 2012. Approaches for Sparse Unsupervised Learning. In: Proceedings of the 29th International Conference on Machine Learning (ICML-12), pp: 751-758.
15. Nair, V. and G.E. Hinton, 2009. 3D object recognition with deep belief nets. In: Advances in Neural Information Processing Systems, pp: 1339-1347.
16. Sutskever, I., Martens, J. Dahl and G. Hinton, 2013. On the importance of initialization and momentum in deep learning. ICML.
17. Štrumbelj, E., Z. Bosni´c, I. Kononenko, B. Zakotnik and C. Grašic Kuhar, 2010. Explanation and reliability of prediction models: The case of breast cancer recurrence. Knowledge and Information Systems, 24: 305-324.
18. Tomczak, J.M., 2013. Prediction of breast cancer recurrence using Classification Restricted Boltzmann Machine with Dropping. arXiv preprint arXiv:1308.6324.
19. Tomczak, J.M. and A. Gonczarek, 2013. Decision rules extraction from data stream in the presence of changing context for diabetes treatment. Knowledge and Information Systems, 34 (3): 521-546.
20. Zhou, S.K. and R. Chellappa, 2006. From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel hilbert space. IEEE Transactions on Pattern Analysis and Machine Intelligence. 28 (6): 917-929.
21. Zieba, M., J.M. Tomczak, M. Lubicz, J. S´wia?tek, 2014. Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. Applied Soft Computing, 14 (A): 99-108.