A study on using deep autoencoders for imbalanced binary classification

Vlad-Ioan Tomescu, Gabriela Czibula, Ștefan Nițică

Department of Computer Science, Babeș-Bolyai University

1, M. Kogalniceanu Street, 400084, Cluj-Napoca, Romania

25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2021)

Outline

Plan

Introduction

- Imbalanced classification represents a challenge for supervised learning, due to poor predictions on the minority class.
- Classifiers tend to mostly predict the majority class.
- Autoencoders, traditional feature extractors, used for binary classification.
- Application field = Medicine, more specifically breast cancer detection.

Plan

Background

- According to the World Health Organisation (WHO), **breast cancer** (BC) is the most frequent form of cancer among women
- It is responsible for 15% of all cancer-related deaths in this group, with 627,000 cases reported only in 2018.
- The currently used screening methods are not able to detect breast cancer at its earliest stages, when the chances of saving the patients' lives are maximal [Org19]
- Mammography, which is the main breast cancer screening procedure employed worldwide, has proven to have no significant impact in reducing the mortality rate

Related work

- Ohja and Goel [OG17] analysed the performance of different clustering and supervised classification algorithms.
- Borges [Bor15] compared two ML techniques (Bayesian Networks and J48) for BC classification and applied them on Wisconsin Breast Cancer Diagnosis data set. [WSM]
- Kumar et al. [KMM+20] comparatively applied twelve classification techniques in predicting breast cancer.
- Rehman et al. [RZMA+19] performed a study on breast cancer detection by developing Random Forest (RF) and Support Vector Machine (SVM).
- Cervo et al. [RZMA+19] applied PCA-LDA analysis on SER spectra acquired on blood serum samples.

Plan

t-SNE

- **t-SNE** [vdMH08] is a nonlinear dimensionality reduction technique
- It belongs to unsupervised learning
- Maps a probability distribution from a high dimensional input space into a lower dimensional space
- Aims at maximizing the similarity between distributions (unlike other dinesionality reduction techniques). It uses the Kullback-Leibler (KL) divergence for that.

Autoencoders

- An **Autoencoder** (AE) [GBC16] is a self-supervised feed forward neural network
- Aims to learn the identity function, more specifically to recreate the input
- Has two main components:
  - The **encoder**, which maps the *n*-dimensional input space into an *m*-dimensional hidden space
  - The **decoder**, which learns to reconstruct the original input space from the hidden space
- Can be used for dimensionality reduction, if the dimentionality of the hidden space is lower than the one of the input space

Support Vector Machines

- **Support vector machines** (SVMs) are supervised learning methods used for both classification and regression [PS20].
- SVMs separate hyperplanes in high dimensional space, given a set of high dimensional real valued vectors (data points).
- The optimisation problem consists of maximizing the distance of each class from the hyperplane.

Plan

Data sets

- Three data sets will be used in our study. Each data set consists of samples corresponding to BC patients belonging to two classes: *benign* or *malignant*.

Data set	Acronym	# of attributes	Positive instance	Negative instance	Classes
Wisconsin Breast Cancer (Original) [25]	WBC	9	239	444	"D''_+ = *malignant* "D''_- = *benign*
Wisconsin Diagnostic Breast Cancer [27]	WDBC	30	212	357	"D''_+ = *malignant* "D''_- = *benign*
SERS data set [4]	SERS	1321	40	20	"D''_+ = *BC* "D''_- = *healthy*

Figure: Description of the BC data sets used.

Data sets

- The WBC [Wol] data set is considered the easiest, the WDBD [WSM] is of average difficulty, while SERS [CMM+15] presents the highest challenge.
- The difficulty of SERS arises from the low number of instances and a lack of a clear separation between classes, based on the the value of those instances.
- Sers is also the only data set where there are (considerably) more positive instances than negative ones.
- All 3 data sets are imbalanced.

t-SNE representations of data sets

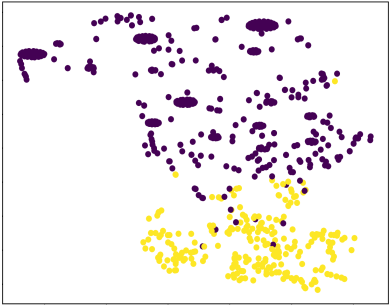

Figure: 2D t-SNE visualisation of WBC. The *benign* class is coloured with purple and the *malignant* class with yellow.

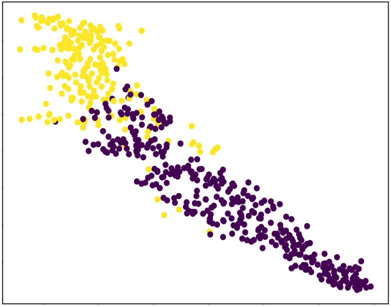

Figure: 2D t-SNE visualisation of WDBC. The *benign* class is coloured with purple and the *malignant* class with yellow.

t-SNE representations of data sets

- All 3 models show a clear separation between classes, with some outliers present.

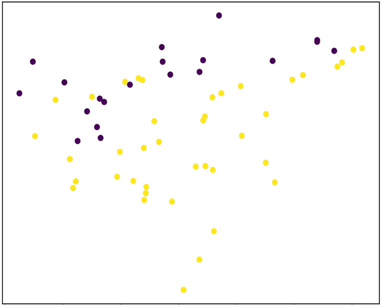

Figure: 2D t-SNE visualisation of SERS. The *benign* class is coloured with purple and the *malignant* class with yellow.

Plan

Methodology

- Autoencoders are traditionally used for feature extraction in unsupervised learning.
- We are also using them for supervised learning, namely binary classification.
- Each autoencoder used was trained on only 1 class, resulting in smaller loss on that class.
- The data points of each class were split into Train, Val, Test
- Cross-validation over 10 iterations

Methodology

Three experiments were conducted:

- Training 1 Autoencoder on the majority class. Prediction done by setting a dynamic threshold, so that the performance on both classes is balanced.
- Training 2 Autoencoders, each on a different class. Prediction done by choosing the autoencoder with the smaller loss.
- Training 2 Autoencoders as before. The pair of losses considered 2d points and classified using SVM.

Plan

Results

Data set	Model	Performance estimation	Acc	PPV	NPV	$Sens$	$Spec$	AUC	F-$score$
WBC	C_{1AE}	**Best**	0.986	0.962	1.000	1.000	0.978	0.989	0.985
		Overall	0.944± 0.02	0.907± 0.04	0.967± 0.02	0.940± 0.03	0.946± 0.02	0.943± 0.02	0.939± 0.02
	C_{2AE}	**Best**	0.972	0.960	0.978	0.960	0.978	0.969	0.972
		Overall	0.915± 0.03	0.956± 0.04	0.903± 0.03	0.800± 0.07	0.978± 0.02	0.889± 0.04	0.913± 0.03
	$C_{2AE-SVM}$	**Best**	0.986	0.962	1.000	1.000	0.978	0.989	0.986
		Overall	0.961± 0.02	0.926± 0.03	0.983± 0.02	0.968± 0.03	0.957± 0.02	0.962± 0.02	0.961± 0.02
WDBC	C_{1AE}	**Best**	0.966	0.917	1.000	1.000	0.944	0.972	0.964
		Overall	0.878± 0.03	0.807± 0.04	0.933± 0.03	0.895± 0.05	0.867± 0.03	0.881± 0.03	0.872± 0.03
	C_{2AE}	**Best**	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		Overall	0.945± 0.02	0.977± 0.03	0.931± 0.03	0.878± 0.06	0.986± 0.02	0.932± 0.03	0.944± 0.02
	$C_{2AE-SVM}$	**Best**	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		Overall	0.945± 0.02	0.987± 0.02	0.928± 0.04	0.868± 0.07	0.992± 0.01	0.930± 0.03	0.943± 0.02
SERS	C_{1AE}	**Best**	0.833	1.000	0.750	0.875	1.000	0.813	0.813
		Overall	0.685± 0.08	0.870± 0.07	0.532± 0.10	0.651± 0.13	0.750± 0.15	0.701± 0.07	0.662± 0.08
	C_{2AE}	**Best**	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		Overall	0.850± 0.07	0.988± 0.02	0.730± 0.09	0.788± 0.11	0.975± 0.05	0.881± 0.06	0.851± 0.07
	$C_{2AE-SVM}$	**Best**	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		Overall	0.758± 0.08	0.950± 0.07	0.609± 0.10	0.675± 0.10	0.925± 0.10	0.800± 0.08	0.761± 0.08

Figure: Experimental results.

Results

- For larger and easier data sets, $C_{2AE-SVM}$ is the best model, due to the extra learning provided by the SVM.
- On harder data sets, the SVM tends to overfit
- Performance on each class, similar for C_{1AE}. Sensitivity close to specificity. This can be adjusted by moving the decision boundary.

Results

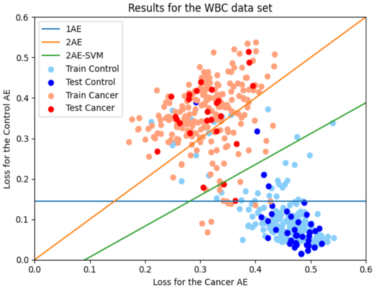

Figure: Losses for the WBC data set together with the decision boundaries generated. The OX axis represents the loss values for the "+" class, while the OY axis expresses the loss values for the "-" class

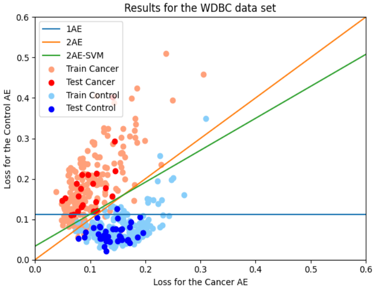

Figure: Losses for the WDBC data set together with the decision boundaries generated. The OX axis represents the loss values for the "+" class, while the OY axis expresses the loss values for the "-" class

Results

- The decision boundaries vary with the nature of the data set.

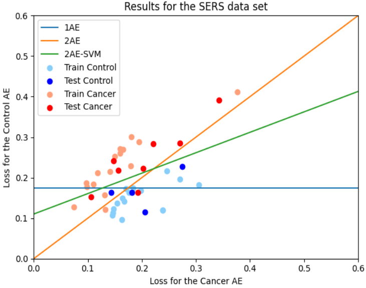

Figure: Losses for the SERS data set together with the decision boundaries generated. The OX axis represents the loss values for the "+" class, while the OY axis expresses the loss values for the "-" class

Comparison to related work

- Our models can be compared to the literature.
- These models from literature also employ cross-validation.

Data set	**Our model**		LDA	GNBC	DT	MLP	SVC	LR	kNN	AB	RF	SGD
WBC	$C_{2AE-SVM}$	0.962 ± 0.02	0.959 ± 0.02	0.973 ± 0.01	0.934 ± 0.01	0.963 ± 0.01	0.977 ± 0.01	0.971± 0.01	0.978± 0.01	0.961± 0.02	0.972± 0.01	0.967± 0.01
WDBC	C_{2AE}	0.932 ± 0.03	0.951± 0.01	0.933± 0.01	0.920± 0.02	0.920± 0.02	0.894± 0.02	0.929± 0.02	0.919± 0.01	0.953± 0.01	0.952± 0.01	0.858± 0.05
SERS	C_{2AE}	0.881± 0.06	0.767± 0.08	0.854± 0.06	0.704± 0.10	0.850± 0.06	0.601± 0.10	0.500± 0.00	0.734± 0.11	0.792± 0.10	0.833± 0.08	0.696± 0.11

Figure: AUC values for our best performing model and eight classifications models: LDA, GNBC,DT, MLP, SVC, LR, kNN, AB, RF and SGD

Comparison to related work

- Shannon entropy [GK17] represents a measure of imbalancement degree.
- Most significant original contribution done by our $C_{2AE-SVM}$ model on SERS.

Data set	Entropy	Our best model	Win	Lose
WBC	0.934	$C_{2AE-SVM}$	3	7
WDBC	0.951	C_{2AE}	6	4
SERS	0.915	C_{2AE}	10	0

Figure: Summary of the comparison.

Plan

Conclusions

- We introduced three classification models based on deep autoencoders.
- Goal = improve performance on the minority class, in imbalanced classification tasks.
- The applied field of medicine, more specifically breast cancer detection, was chosen, due to the major importance of the problem.
- The relative performance of each model depends on the nature of the data set.

Thank you!
Questions?

Bibliography

Lucas Borges.
Analysis of the wisconsin breast cancer dataset and machine learning for breast cancer detection.
pages 15–19, 10 2015.

Silvia Cervo, Elena Mansutti, Greta Mistro, Riccardo Spizzo, Alfonso Colombatti, Agostino Steffan, Valter Sergo, and Alois Bonifacio.
Sers analysis of serum for detection of early and locally advanced breast cancer.
Analytical and Bioanalytical Chemistry, 407:7503–7509, 2015.

I. Goodfellow, Y. Bengio, and A. Courville.
Deep Learning.
MIT Press, 2016.

Diego Galar and Uday Kumar.

Chapter 3 - preprocessing and features.
In Diego Galar and Uday Kumar, editors, *eMaintenance*, pages 129–177. Academic Press, 2017.

Vivek Kumar, Brojo Kishore Mishra, Manuel Mazzara, Dang N. H. Thanh, and Abhishek Verma.
Prediction of malignant & benign breast cancer: A data mining approach in healthcare applications.
In *Advances in Data Science and Management*, pages 435–442, Singapore, 2020. Springer Singapore.

U. Ojha and S. Goel.
A study on prediction of breast cancer recurrence using data mining techniques.
In *2017 7th International Conference on Cloud Computing, Data Science Engineering - Confluence*, pages 527–530, Jan 2017.

World Health Orgnisation.
Breast cancer: Early diagnosis and screening, 2019.

Derek A. Pisner and David M. Schnyer.

Chapter 6 - support vector machine.
In Andrea Mechelli and Sandra Vieira, editors, *Machine Learning*, pages 101–121. Academic Press, 2020.

Oneeb Rehman, Hanqi Zhuang, Ali Muhamed Ali, Ali Ibrahim, and Zhongwei Li.
Validation of mirnas as breast cancer biomarkers with a machine learning approach.
Cancers, 11(3):431, 2019.

Laurens van der Maaten and Geoffrey Hinton.
Visualizing data using t-sne.
Journal of Machine Learning Research, 9:2579–2605, 2008.

William H. Wolberg.
Breast cancer wisconsin (original) data set.

William H. Wolberg, W. Nick Street, and Olvi L. Mangasarian.
Breast cancer wisconsin (diagnostic) data set.