Learning Vector Quantization with Applications in Neuroimaging and Biomedicine

van Veen, Rick

DOI:
10.33612/diss.211419033

Chapter 1

General Introduction

Machine learning has gained increasing relevance within many medical subdomains [6]. This thesis presents machine learning applications, emphasizing the use of learning vector quantization (LVQ) in neuroimaging and biomedicine.

LVQ is a family of algorithms [7] introduced by Kohonen [8] that construct typical data patterns, called prototypes, representing the classes in a dataset. Predictions are made based on the distance between a novel data sample and these prototypes. LVQ has been applied successfully in the biomedical field, medicine, and industry[1]. Advanced versions, such as generalized matrix learning vector quantization (GMLVQ) [9], generate a relevance matrix to infer which features are most relevant for the classification. A valuable property of these methods is that the prototypes and relevance matrices can, in a typical case, be interpreted directly to gain more insight into the model and the data.

As an example of a typical application of GMLVQ in biomedicine, we discuss the classification of adrenocortical carcinoma (ACC). ACC is a rare disease in which malignant (cancer) cells form in the outer layer of the adrenal gland. A non-invasive, accurate, and inexpensive test to differentiate ACC from other adrenal tumors is urgently needed. We study steroid measurements from patients with malignant and benign tumors measured using two different measuring techniques. Using GMLVQ, we set out to solve the classification problem and, in addition, decide which steroids are most relevant with the goal of choosing the optimal measuring technique.

The main focus is on the application of LVQ for the classification and interpretation of neuroimaging data collected from patients with different neurodegenerative diseases. Neurodegenerative diseases are incurable, characterized by progressive loss of function of nerve cells in specific brain regions or peripheral nervous systems, ultimately resulting in the patient's death. The World Health Organization has suggested that the motor function affecting neurodegenerative disorders can be the second-most prevalent cause of death after cardiovascular diseases [10, 11]. Unfortunately, neurodegenerative disorders can be complex to diagnose, and no diagnostic test exists to differentiate between them in vivo. The diagnosis, therefore, currently relies on the experience of the treating physician, resulting in unreliable di-

[1] http://www.cis.hut.fi/research/som-bibl/

agnosis at an early stage of the disease. Neuroimaging techniques have significantly developed in recent decades [12] and might prove effective in overcoming this problem. One such method is positron emission tomography (PET), a scanning technology from nuclear medicine that measures a tracer's distribution and concentration [13]. This specific scanning technology combined with the [^{18}F] fluorodeoxyglucose (FDG) tracer can quantify functional changes in a patient's brain and aid the early diagnosis of degenerative diseases.

We study the performance of LVQ for the classification of neurodegenerative diseases using data collected at single and multiple neuroimaging centers. In addition, we present a novel method to deal with issues encountered with multi-source data.

1.1 Scope of this Thesis

This thesis contains work on the applications of machine learning methods within the domain of biomedicine and neuroimaging. The main focus is on the application of LVQ in neuroimaging, with as goal to classify neurodegenerative diseases.

The biomedicine study (Section 3.2) provides a typical example of the application of GMLVQ, where one can directly understand the prototypes and relevance matrix. The goals of this study are: first, to build a system with good diagnostic performance. Second, to compare these systems when trained on two different (throughput and cost) steroid measuring technologies.

We need to deal with limited amounts, high-dimensional, and potentially mislabeled data to diagnose neurodegenerative disease using neuroimaging data. This thesis looks into resting-state FDG-PET data and pre-processes these with the scaled subprofile model (SSM)/principal component analysis (PCA) method. As this process complicates the direct interpretation of the feature space and LVQ models, we present a method to revert the process and regain interpretability. A considerable amount of effort has been put into analyzing the data, as seen by the GMLVQ systems, to point out potential exceptional subjects. We study the classification of neurodegenerative disease using data collected at multiple neuroimaging centers. We investigate the possibility of a universal classifier within the context of this dataset. Additionally, we introduce and investigate a novel method of dealing with the issues encountered during this investigation.

Lastly, from a technical perspective, we present our implementation of generalized learning vector quantization (GLVQ), GMLVQ, and localized generalized matrix learning vector quantization (LGMLVQ) (Chapter 3). Which we provide as an open-source, general, expandable, well-documented, and tested LVQ Python package[2].

1.2 Outline

The following chapter will define and discuss the domain and technical concepts relevant to the proceeding chapters. Chapter 3 presents "sklvq," our Python im-

[2] `https://github.com/rickvanveen/sklvq/`

plementation of LVQ, addressing some of the issues with current toolboxes. Additionally, in this chapter, we present a typical application of GMLVQ in biomedicine. Chapter 4 will show how one can use GMLVQ to analyze a neuroimaging dataset to improve the currently available diagnoses. Machine learning algorithms generally need a lot of data, and a way to achieve this is to combine data from multiple sources. Chapter 5 will present the study that shows the problems encountered classifying neurodegenerative diseases using data from three neuroimaging centers. In addition, we present the combination of SSM/PCA features and LVQ prototypes and relevance matrix in the original voxel space. A novel and generally applicable extension to the training procedure of GMLVQ to correct for multi-source data variation will be presented in Section 6.1.1. Finally, we will give a summary of our findings and outlook in Chapter 7.