# Abstract

Nearest Neighbour (NN) classifiers are one of the most celebrated algorithms in machine learning. In recent years, interest in these methods has flourished again in several fields (including statistics, machine learning and pattern recognition) since, in spite of their simplicity, they reveal as powerful non-parametric classification systems in real-world problems. The present work is mainly devoted to the development of new learning algorithms for these classifiers and is focused on the following topics:

- ❑ Development of learning algorithms for crisp and soft k-NN classifiers with large margin
- ❑ Extension and generalization of Kohonen’s LVQ algorithms
- ❑ Local stabilization techniques for ensembles of NN classifiers
- ❑ Study of the finite-sample convergence of the on-line LVQ1 and k-means algorithms

Besides, a novel oriented principal component analysis (OPCA) addressed for feature extraction in classification is introduced. The method integrates the feature extraction into the classifier and performs global training to extract those features useful for the classifier. The application of this general technique in the context of NN classifiers derives in a problem of learning their weight metric.

**Index Terms-** Soft k-Nearest-Neighbour Classifiers, Nearest-Neighbour Classifiers, Large Margin Classifiers, Oriented Principal Component Analysis, Local Stabilization, Ensemble Learning, Batch Learning Vector Quantization (LVQ) algorithms, Generalised LVQ1, Batch LVQ1, Dynamic LVQ algorithms, Kohonen’s LVQ algorithms, Finite-sample Convergence, Online gradient descent, Hand-written Character Recognition.

# Index

## PART I: Introduction

## PART II: Learning Algorithms based on Supervised Clustering for Nearest Neighbour Classifiers

## PART III: Learning Algorithms for Large Margin Nearest Neighbour Classifiers

## PART IV: Global training algorithms for Pattern Recognition

## PART V: Conclusion

# 1 Introduction

Ill-defined Engineering Problems

Engineering must deal with problems that are *ill defined* so there is no mathematical theory that can properly define them. Traditionally, engineers have coped with this kind of problems designing ad-hoc strategies based mainly on a corpus of heuristics collected through decades of research. However, there are more systematic ways of building machines that solve these difficult real-world problems (e.g. handwriting recognition).

Statistical Learning Machines

If we can extract from these problems measures (or examples) that reflect their behaviour, we could build, with the help of a **learning machine**, a model or a device that, given certain conditions, could reflect the computational structure of the problem. What is typically inferred from this data is a statistical model that deals with the inherent uncertainty or imprecision of the examples.

Capacity Control

This model has a set of adjustable parameters that are estimated in the learning phase using a set of examples (the training set). Nevertheless, the learning machine must *control its capacity* (e.g. the effective number of free parameters) to ensure a reliable estimation of their parameters and consequently good *generalization* (e.g. correct response to unseen examples). Hence, the learning device must solve efficiently a trade-off between its capacity and the information about the problem given by the training set. This problem can also be formulated as a bias-variance trade-off or a balance between the approximation error and the estimation error of the learning machine. However, all these formulations are qualitatively similar.

Ensemble Learning

Many recent efforts in machine learning address the problem of **increasing generalization through capacity control**. One example is **ensemble learning**. The idea of ensemble learning is to combine an uncorrelated collection of learning systems that have been all trained in the same task. This combination is typically done by majority in classification or by averaging in regression. Generally speaking, these techniques controls capacity stabilising the solution through the reduction of dependence on the training set and the optimization algorithms used by the members of the ensemble.

Large-Margin Classifiers

Another examples of methods for controlling capacity are **large margin classifiers**. In classification problems, the learning machine must assign input patterns to one of the pre-defined categories. Typically, these systems are designed to minimise the number of misclassifications in the training set. However, recently it has been showed that, in order to ensure a small generalization error, we should also take into account the *confidence* of the classifications. Then classifiers must also be designed to have a *large margin distribution* of the training samples; that is, the training samples must be assigned to the correct class with high confidence. A large margin distribution helps to stabilise better the solution and hence capacity can be controlled. Two examples of large margin classifiers are *support vector learning machines* (SVM) and *boosting classifiers*.

Large-scale Learning

Other recent efforts address the problem of **scaling up the learning algorithms** for handling difficult high-dimensional real-world problems with large databases. Current developments in this area include modular and hierarchical networks, and other forms of *co-operative learning machines*. Many of the existing approaches use *gradient-based learning* as the unifying principle for training the whole system in a global way. Then they back-propagate errors through their complex architectures to compute the update equations of a **global training** algorithm.

Scope of the thesis

Pattern Recognition

This thesis addresses the problem of **learning pattern recognition** in the light of current developments in machine learning. The goal in pattern recognition is to synthesise reliable machines that group complex input data (or patterns) into categories (or classes) with the help of a (supervised) learning device that uses a set of labelled patterns. The core of a pattern recogniser is usually composed of a feature extractor and a classifier. The feature extractor reduces the input by measuring certain invariant "features" or "properties". (This helps to reduce the complexity of the original problem and the design of the classifier. Consequently, feature extraction is also a mechanism to control the capacity of the recogniser.) The classifier uses then these features to make the decision of assigning the input pattern to a class.

Automatic Feature Extraction and Selection

The **feature extractor** was often handcrafted since it is rather specific to the problem. However, the current tendency is rely more on learning devices that *automatically extract features* and less on manual feature extraction of discriminatory information. Unsupervised learning algorithms are commonly used to build feature extractors from training data. However the application of these algorithms can lead to lose important discriminatory information since they do not take into account class labels.

Global Training

Oriented Principal Component Analysis

An alternative and powerful implementation is to *integrate the feature extractor into the classifier* and to perform a global training of both systems to alleviate the problem of separate and uncoupled training. The thesis proposes a novel method (called *oriented principal component analysis,* OPCA) to perform a *global gradient-based training* of a feature extractor that uses several lineal combinations of input variables and any classifier that allow a back-propagation of an error signal through its architecture (e.g. feed-forward networks).

There are many **classification methods** but among them, **nearest neighbour (NN) classifiers** are one of the most famous in machine learning. Researches have dedicated new

Nearest Neighbour Classifiers

attention to these methods in several fields like statistics, machine learning and pattern recognition since these lazy learning or memory-based methods reveal, in spite of their simplicity, as very powerful non-parametric classification systems in real-world problems. When an input pattern is presented to these classifiers, they compute the k closest prototypes to it using a distance metric defined by the user. Then the classifier assigns the class label using a majority vote among the labels of the k nearest prototypes. (If k=1 the classifier simply assigns the label of the nearest prototype to the input pattern.) The parameters of these classifiers to be estimated in the learning phase are often the set of prototypes and (sometimes) the distance metric. (E.g. OPCA in the context of NN classifiers derives to a problem of learning their distance metric.) The most direct approach to compute the prototypes is to store the whole training database as the set of prototypes. However, storage and computational requirements and the belief that simpler solutions achieve better generalization (Occam's razor) advocate the use of more condensed sets.

Condensed Nearest Neighbour Classifiers

Soft k-NN estimation

The thesis is also devoted to the development of new learning algorithms that compute a reduced number of prototypes for NN classifiers. Our work also proposes a simple extension of crisp k-NN classification called *soft k-NN* that uses a locally defined Parzen window using the k-nearest prototypes to the input pattern.  Soft k-NN improves the approximating capabilities of the posterior class probabilities estimates of its crisp counterpart while it can give soft outputs for post-processing purposes.

Large Margin Nearest Neighbour Classifiers

The thesis develops new learning algorithms for crisp and soft k-NN classification that minimises the training error and also achieves a large margin distribution of the training samples. The resulting classifiers are accordingly called *large margin nearest neighbour classifiers*. Interesting relationships between large margin NN classifiers and SVM are pointed out in our work.

Local stabilisation of Ensembles of NN classifiers

Ensemble learning in NN classifiers is also addressed. We derive two methods (called *local averaging* and *local extreme*) that give a single predictor based on a stabilisation of the classifiers. However our methods for ensemble learning differ conceptually from the others in two main points. First, our methods take advantage of the local nature of the NN classifiers *stabilising locally* (inside the classifier) while existing ensemble methods fuse at output level. Besides, the idea of our methods is to obtain a single predictor that outperforms the best member of the ensemble but having the same complexity (e.g. the same number of prototypes of each single member). On the other hand, the other methods for ensembles compute a predictor that uses all the members of the ensemble (e.g. bagging). This second feature makes our ensemble methods suitable for real-world applications since the other techniques are often prohibitive from a practical point of view due to they use typically hundreds or thousands of individual classifiers to build a single predictor

LVQ1 based on Newton

GLVQ1

BLVQ

DLVQ

Extension and generalization of techniques based on supervised clustering like the powerful Kohonen's LVQ algorithms are also proposed. We present a second order version of the LVQ1 algorithm based on Newton optimization to speedup the convergence rate and a simple generalization of LVQ1 that improves the probability of reaching a minimum of the training error. We also derive a family of batch LVQ algorithms using the idea of supervised clustering in a more principled way. We perform a clustering process over a modified probability density function which is zero at Bayes borders obtaining a very fast learning algorithm for Euclidean NN classifiers (i.e. NN classifiers that use the Euclidean distance as a metric). The problem of determining proper initial values for these algorithms and the optimal number of prototypes is studied and a dynamic LVQ algorithm is proposed.

Finite-sample Convergence of on-line algorithms

As we have already seen, gradient-based learning is the core of many learning algorithms that deal with practical applications. Typically, the on-line (or stochastic) version (in which the iterative equation of the learning algorithm are updated after each pattern presentation) is preferred since they can converge faster than batch versions and also their dynamics are noisier so they could escape from local minima points of the cost function. Online gradient-based learning are very simple optimization algorithms so they are not powerful to find any feasible solution. However, the use of more precise optimization algorithms (e.g. second-order methods like Newton) are often discarded in real-world applications due to their excessive computation requirements. Besides, the use of a better optimization algorithm does not guarantee better generalization performance of the learning machine. As it has been observed, the *under-computation* of gradient system prevents them to avoid over-fitting (i.e. an excessive tuning to the training set). The implicit limitation of the hypothesis space (i.e. the space where the learning algorithm finds a solution) of the gradient systems is therefore an effective way of controlling the capacity of the learning machine. Hence, the study of on-line gradient algorithms is of central interest in machine learning. More precisely, the analysis of the conditions to ensure the convergence of these algorithms to the desired solution is fundamental. Large-sample asymptotic convergence of on-line gradient algorithms (i.e. convergence near an attraction basin when the number of training samples tends to infinite) have already been studied using tools of the stochastic approximation theory. However, these tools are (in principle) useless for the finite-sample case. Finally, we address a novel study of the finite-sample convergence (i.e. the real convergence for finite training sets) of two learning algorithms for the design of NN classifiers: the on-line LVQ1 and k-means algorithms. This new study is based on the use of tools of the dynamic system and optimization theories. We compute the real attractors (i.e.. the equilibrium points) of the learning equations and the conditions to ensure good convergence and low optimization error.

## 1. Scope of the Thesis

The main goal of this dissertation is the development of new and powerful learning algorithms for crisp and soft k-nearest-neighbour classifiers in the light of recent advances in machine learning.

For this aspiration we start with the study of some existing approaches based on clustering like the k-means and Kohonen's LVQ1 algorithms. We analyse the finite-sample convergence (that is the real convergence) of these algorithms. Then we extent and generalise Kohonen's LVQ algorithms.

First, we simply derive the LVQ1 algorithm using the Newton optimization and extent it to a generalised LVQ1 (GLVQ1) algorithm (a constrained extension of the LVQ1) that includes the LVQ1 and k-means.

GLVQ1 is a powerful (though very simple) extension but it has the drawback that its regularising parameter must be empirically determined. A family of batch LVQ algorithm that employs Newton optimization is then presented. This family of algorithms is derived using a more principled approach based on the idea of performing a clustering process over a modified probability density function which is zero at Bayes borders.

The problem of determining proper initial values for these algorithms and the optimal number of prototypes is studied and a dynamic LVQ algorithm is proposed.

Finally, the issue of stabilising an ensemble of NN classifiers trained with these algorithms is addressed and two methods based on the local nature of NN classifiers are proposed to achieve a single predictor with a similar number of parameters than the members of the ensemble.

Then we address the derivation of new learning algorithms for crisp and soft k-NN classification than minimises the training error and also achieves a large margin distribution of the training samples. The resulting classifiers are accordingly called *large margin nearest neighbour classifiers*. Interesting relationships between large margin NN classifiers and SVM are pointed out in our work.

Finally, the thesis proposes a novel method (called *oriented principal component analysis,* OPCA) to perform *global gradient-based training* of a feature extractor that uses several lineal combinations of input variables and any classifier that allow a back-propagation of an error signal through its architecture (e.g. feed-forward networks). The idea of the method is to perform a series of training sessions until it finds an optimal projection of the input variables in the feature space that allows a better separation of classes.

## 2. Contributions of the Thesis

We have obtained the following results, listed according to their relative importance:

- ❑ Development of learning algorithms for crisp and soft k-NN classifiers that have a large margin distribution of training patterns while minimises the training error. Relationships between *large margin NN classifiers* and SVM are pointed out in our work. (Chapters 8 and 9)
- ❑ Elaboration of a novel *oriented principal component analysis* (OPCA) addressed for finding optimal feature extraction in classification problems based on a global gradient training of the feature extractor and the classifier. The proposed method can use any classifier that allow the back-propagation of an error signal through its architecture (e.g. feed-forward networks). (Chapter 10)
- ❑ Development of *local stabilization techniques for ensembles of NN classifiers*. The two new methods addressed for NN ensembles use local stabilisation techniques to derive a single predictor that have the same complexity than each single member of the ensemble. Consequently, our methods differ from current ensemble techniques that work at a global level and use the whole ensemble in the stabilised predictor. (Chapter 7)
- ❑ Extension and generalization of Kohonen’s LVQ algorithms. These include: 1) a second order version of the LVQ1 algorithm based on Newton optimization to speedup the convergence rate; 2) a simple generalization of LVQ1 that improves the probability of reaching a minimum of the training error; 3) a family of batch LVQ algorithms using the idea of supervised clustering of Kohonen’s LVQ algorithms in a more principled way and 4) a dynamic LVQ algorithm to mitigate in some degree the problem of determining proper initial values for these algorithms and the optimal number of prototypes. (Chapters 4, 5 and 6)
- ❑ Study of the finite-sample convergence of the on-line LVQ1 and k-means algorithms. This novel study uses tools of the dynamic system and optimization theories for computing the real attractors (i.e. the equilibrium points) of the on-line learning equations and the conditions for ensuring good convergence and low optimization error. (Chapters 3 and 4)

All the programs to implement the learning algorithms and other functions to test different statistics of the NN classifiers have been developed using the C programming language. More than 150000 lines of code have been elaborated. The LVQ_PAK library has been used as the core library of our own libraries. The LVQ_PAK package can be found at http://www.hut.fi/.

## 3. Structure of the Thesis

The thesis contains original work on nearest neighbour methods and feature extraction for classification and is structured in four main parts and several sub-parts:

- ❑ Introductory part (this chapter and chapter 2)
- ❑ Learning algorithms based on clustering (chapters 3-7)
  - ❑ Study of finite-sample convergence of k-means (chapter 3) and LVQ1 (chapter 4)
  - ❑ Improvement in the optimization algorithm of LVQ1 and its generalization (GLVQ1) (chapter 4)
  - ❑ Generalization of Kohonen's LVQ algorithms (BLVQ) (chapter 5)
  - ❑ Constructive and Pruning algorithms for LVQ algorithms (DLVQ) (chapter 6)
  - ❑ Local stabilization of LVQ-based NN ensembles (local averaging and local extreme) (chapter 7)
- ❑ Learning algorithms for large margin crisp and soft k-NN classifiers (chapters 8 and 9)
- ❑ Optimal feature extraction for classification based on OPCA (chapter 10)
- ❑ Conclusions (chapter 11)

All the chapters are self-contained and cross-references between chapters have been minimised so the reader can start at any place. However, the following roadmap is recommended:

In a first reading, one can read chapters 2 and 11. Chapter 2 gives a brief introduction to the work developed in this thesis while chapter 11 presents some conclusions; then the reader can go through several chapters depending on his/her interests.

Finite-sample convergence properties of on-line algorithms are first introduced in chapter 3 (on-line k-means) and then in chapter 4 (on-line LVQ1). Since both on-line learning systems have similar behaviours is recommended to read chapter 3 and then chapter 4.

Learning with supervised clustering algorithms for NN classifiers is studied in chapters 3-7. The study of LVQ1 and its improvements are presented in chapter 4. A generalisation of the idea of Kohonen's LVQ algorithms based on supervised clustering is introduced in chapter 5

(BLVQ). Then chapter 6 proposes constructive and pruning algorithms to dynamically add and delete prototypes during learning using the LVQ algorithms as a core. Finally, the local stabilisation of NN ensembles trained with LVQ algorithms are presented in chapter 7. If the reader is interested in BLVQ, it is recommended to read first chapter 4 to get some flavour of the role of supervised clustering in the design of NN prototypes. On the other hand, the other two chapters can be read independently.

Large-margin crisp and soft k-NN classifiers are introduced in chapters 8 and 9 respectively. Since crisp and soft classifiers are intimately linked the reader could go through chapter 8 first.

Finally, OPCA is introduced in chapter 10 and can be read without the knowledge of the other chapters. Figure 1 shows several routes of reading according to the above considerations.

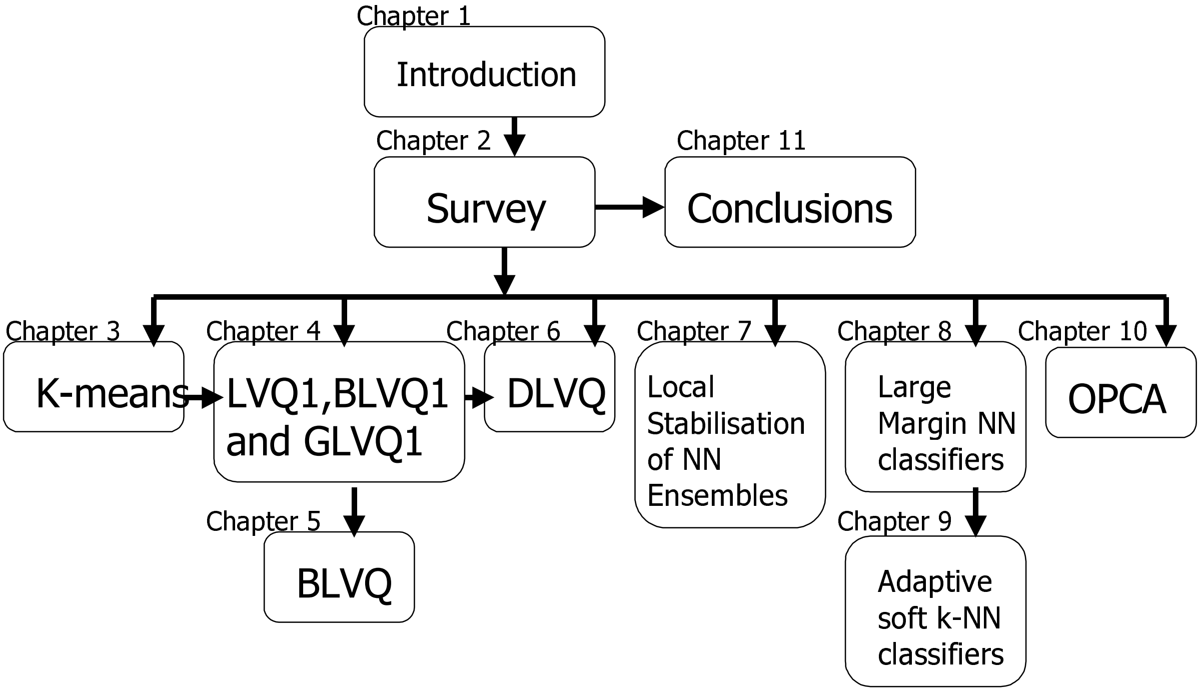

Fig.1.Possible routes of reading.