Chapter 5

Application Area Analysis

In our research regarding ensemble systems, it was necessary to identify different scenarios that allowed us to evaluate the ensemble system's performance against other methods and the different approaches for the preprocess and train stages of the learning process. Such scenarios are represented by different datasets covering different application areas of machine learning like Bioinformatics and Computer Vision. The focus of the experimentations was:

- To compare the general performance of ensemble systems against other methods on benchmark datasets.
- To compare different Train and Preprocess methods and identify the best approach for a certain dataset.
- To apply experimentation results on the design of an Agent System for a problem in bioinformatics for which a solution using classification has not been approached.

The results of the experimentation process will not only bring more confidence on the performance of the framework but will also provide guidelines on the best practices for the design of such ensemble systems.

5.1. Experimentation focus

5.1.1. Benchmark Comparison

In the article **Multi-agent reinforcement learning: weighting and partitioning** [26], Sun arguments that a set of agents will always outperform a single agent, based on the insight that collectiveness helps overcome the deficiencies of a single prediction function. In real life, however, such assumption may be dependent on the dataset and the ensemble method used. In our experiment, 3 benchmark datasets for which results with other techniques have been identified were used to compare the performance of different ensemble techniques. Such datasets are the **Handwritten Digits** [3], the **Secondary Structure Prediction in Proteins** [7] and the **House cost prediction** [12] datasets.

5.1.2. Preprocess Component Analysis

A second focus on experimentation was to evaluate different preprocessing techniques and their effect on the learning results. An special focus was placed on partition and feature reduction. For the feature reduction aspect, it was important to analyze the effect of reducing the number of features in the overall performance. For the partition aspect, the focus was to compare the hard and soft partition procedures and analyze the results. The **Handwritten Digits** [3] dataset was as well used in this section.

5.1.3. Learning algorithm analysis

An additional focus of experimentation were the different ensemble algorithms. First, it was important to analyze which ensemble algorithm showed a better overall performance, so that it should be used on the final experiment. On classification, Adaboost,

Marginal Adaboost and Adaboost* were compared. On prediction, bagging was analyzed. The datasets used in this section were the **Handwritten Digits** [3], the **Secondary Structure Prediction in Proteins** [7] and the **House cost prediction** [12] datasets.

5.1.4. Learning on novel datasets

The results of the previous experiments helped to set the characteristics of the final experiment over a HLA dataset [20], a dataset in the area of bioinformatics which represents a classification problem that is solved using an ensemble multi-classifier in this research. The obtained results will be compared to the results using other approaches, some of them not necessarily in the area of machine learning.

5.2. Datasets Description

In this section, an overview regarding the main characteristics of each dataset is given.

5.2.1. Handwritten Digit Recognition Dataset

This dataset [3] consists on normalized bitmaps of handwritten digits extracted from a preprinted form (Figure 5.1). A total of 43 people contributed in the dataset, being the contribution of 30 used for training and the contribution of 13 used for testing. Each 32x32 bitmap was then divided into 4x4 blocks and the number of pixels inside of each block was then counted, producing a 8x8 input matrix where each element is an integer in the range of 0..16 (Figure 5.2). This procedure was intended for reducing dimensionality and providing invariance to small distortions. The digits as well were

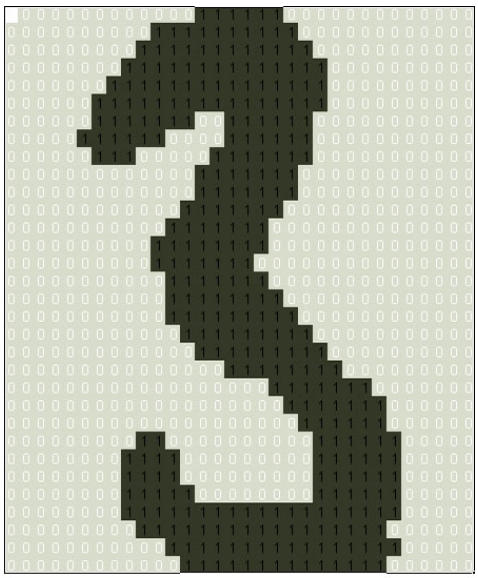

Figure 5.1: Sample input handwritten digit, part of train component of the dataset

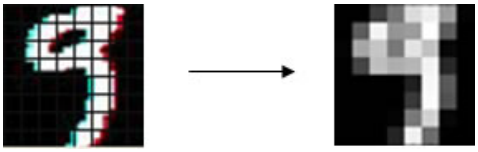

Figure 5.2: Feature simplification of input number. The number of pixels in a 4×4 window are counted, The procedure is focused on reducing sensibility to small variations.

centered and normalized in the recollection process.

In total, there are 3,823 training samples and 1,979 test samples. Labels are provided for the 10 different digits 0..9, so that in all our experiments, we tested the results for each one of the different digits. A sample Sample in the dataset is:

0,1,6,15,12,1,0,0,0,7,16,6,6,10,0,0,0,8,16,2,0,11,2,0,0,5,16,3,0,5,7,0,0,7,13,3,0,8,7,0,0,
4,12,0,1,13,5,0,0,0,14,9,15,9,0,0,0,0,6,14,7,1,0,0,0

Where the first 64 numbers represent the data and the last number (0) represents the associated label. Previous accuracy results with this dataset using k nearest neighbors [3] are shown in table 5.1

k	*Accuracy (%)*
1	98
2	97.38
3	97.83
4	97.61
5	97.89
6	97.77
7	97.66
8	97.66
9	97.72
10	97.55
11	97.89

Table 5.1: k nearest neighbors results. [3]

5.2.2. Secondary Structure Prediction Dataset

This dataset consists on a set of proteins represented by their primary structure as a sequence of amino acids. There exist 20 different amino acids. The focus is to identify the secondary structure of the protein [7], where the secondary structure is defined as the 3-dimensional form of the protein (Figure 5.3). In general, the secondary structure may be of helix, strand or coil type [15]. Kabsch provides the following definition:

> Cooperative secondary structure is recognized as repeats of the elementary hydrogen-bonding patterns turn and bridge. Geometric structure is defined in terms of the concepts torsion and curvature of differential geometry. [15]

The classification of proteins is a classic problem in bioinformatics, and several solutions have been provided which range from statistical to machine learning methods. In the literature, methods which achieve less than 20 % classification error are available [7]. Such methods take advantage of knowledge in protein structure to achieve better results, and it is not the focus of this research to improve such results. It is the focus to use this dataset to analyze multi-agent systems on the binary classification task of identifying proteins with helix second structure from those which do not have it.

The dataset is provided by the University of Dundee [7]. The data contains infor-

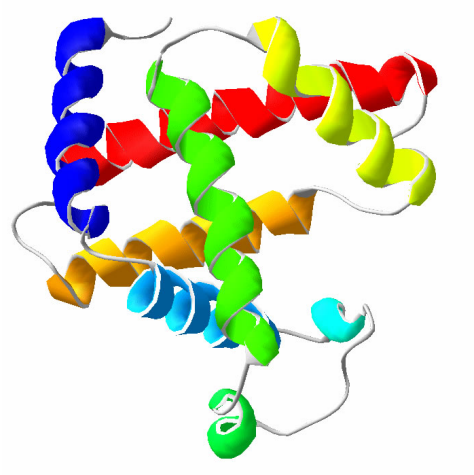

Figure 5.3: Myoglobin protein, presenting helix secondary structure (in color). [2]

mation regarding a 11-sized window from which, each amino acid is represented by a binary vector V, where $|V| = 20$. The vector is 0-valued in every position excepting the position indicating the amino acid. For instance:

$$L(Lysine) = \{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0\}$$

represents the 12th amino acid. The size of the amino acid string is 11, so that there are 297 features. A second set of features was added, following a soft partition approach, indicating the simultaneous presence of 2 amino acids (${}_{297}C_2 = 43956$). In total, there exist 44253 features. The dataset provides Ξ_{train} of size 100 and Ξ_{test} of size 16.

An example sample in the dataset would be [19] [13]

<PROTEIN NAME="1acx", LEN="109">
APAFSVSPASGASDGQSVSVSVAAAGETYYIAQC
APVGGQDACNPATATSFTTDASGAASFSFTVRKS
YAGQTPSGTPVGSVDCATDACNLGAGNSGLNLGH
VALTFG
< /PROTEIN>

5.2.3. HLA Dataset

The HLA Dataset stands apart from previous datasets as this project presents a novel attempt to find a multi-classification system for the solution of the problem.

HLA (Human Leukocyte Antigen) [20] is a genetic marker, which contains the most information regarding histo-compatibility between 2 individuals. Identifying the families to which one individual HLA belongs is crucial in that it indicates whether a certain donor is compatible to a certain transplant recipient.

A HLA sequence belongs to a pair of families F_1, F_2. Each family is represented by a character and a sequence of 4 numbers, for instance **A0123**, however, only the first letter and number are relevant, hence **A01**. In the lab, probes are applied to HLA sequences to discover to which family pair the sequence belongs. Probes react to certain pairs F_i,F_j and do not react to others $F_{i'}$,$F_{j'}$ $i \neq i' \smile j \neq j'$, $F_i, F_j \in F$, where F is the set of all possible families.

Given that we only require the first 2 digits after the group character, we may have a set of families F_i, F_j with the same representation. We may group these families into a group S_k, where $\forall s \in S_k, s = F_i, F_j$. This implies that by uniquely identifying all $s \in S_k$, we uniquely identify S_k and thus identify the family. It is not necessary for us to identify s_i from s_j, but it is important to identify every $s \in S_k$ from every $s \in S_{k'}$

Unique identification using probes

Given a probe p_i, we have information regarding which antigens will present a reaction(which we will represent as 1) and which will not (which we will represent as 0). The original goal regarding HLA is to find the minimal set of probes P, such that every group S is uniquely identified. By unique identification we mean that there is a sequence of probes $p_1, p_2, ..., p_n$ s.t. we are certain that an HLA belongs to a certain

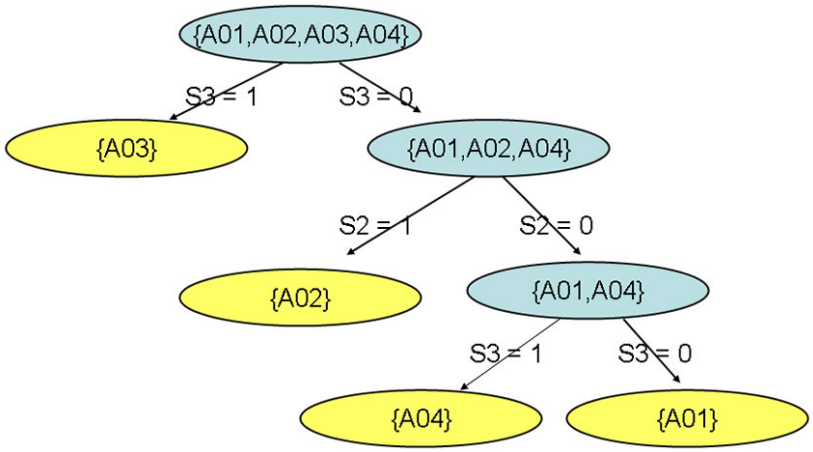

Figure 5.4: Sample binary tree representation of the application of probes for the identification of groups. Groups are separated regarding their rection to different probes (S1,S2,S3), where S = 1 indicates reaction. Leaf nodes represent identified antigens/groups. [4]

family pair.

To model the result of unique identification of groups, a binary tree structure may be used, such that every leaf node will contain only elements which belong to the same group. The minimum set P is such that generates the binary tree with the least depth [4] (Figure 5.4).

Boosting and the HLA problem

As noted before, boosting is based on the concept of having separate weak learners that together will constitute a strong learner. Boosting is a Machine Learning algorithm and to be able to use a machine learning approach, we need to map our problem to a machine learning problem.

Our mapping will be to a classification problem. By classification we refer to successfully identifying the class of a certain object from its contents. Thus, we will rephrase our problem as successfully **classifying HLA to its corresponding group S by its responses to a set of probes**. Table 5.2 shows a sample of group representations by

Group	Family	P_1	P_2	P_3	P_4
1	01	1	0	1	1
1	02	0	1	0	0
1	03	1	0	1	0
2	01	1	1	1	1
2	02	1	0	0	1
3	01	0	1	1	0
3	02	0	1	0	0
3	03	0	0	1	0
3	04	1	1	1	1

Table 5.2: Sample groups with their reaction to probes for the HLA multi-classification problem.

their reactions to probes.

Our Dataset will be defined as follows: We will consider each one of the family pairs as a test sample, and each one of the probes as a feature of it. The result of each feature, 1 or 0, will be a decision stump, where 1 will mean that a certain pair is classified into a certain group and 0 that it is not. As we are mapping this problem to a multi-class classification problem, we will run boosting to obtain the probe set which successfully classifies every element in a certain group S_k as in that group and every element not in that group as not in that group. We will then have a multi-class classifier, with a binary classifier for each group S_k (Figure 5.6). Figure 5.5 shows in general terms the input and output of the HLA multi-classifier.

The problem we are solving will no longer be an optimization problem, however, the classification approach provides advantages in the general solution. Given that each binary classifier is independent, each may be trained separately which contributes to the efficiency of the training process. As well, in the case where the dataset is expanded and a new group is introduced, it can be trained independently without requiring to restart the training process for all other groups. Finally, a prediction on any new antigen may be done by using the resulting multi-classifier.

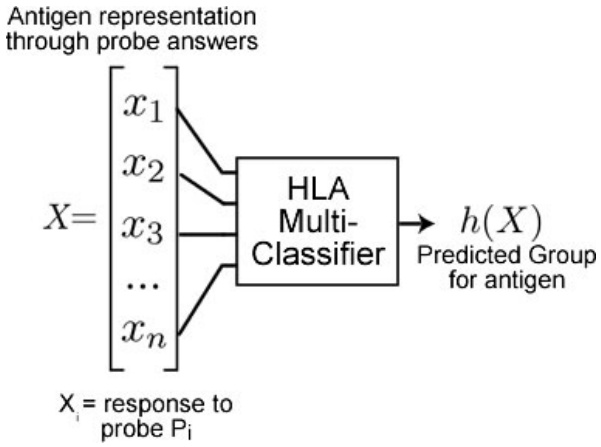

Figure 5.5: Representation for HLA multi-classifier as a black box. Answers to probes are considered as features. The system's response is the predicted family/class of the antigen.

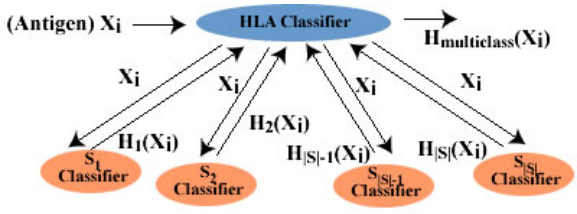

Figure 5.6: Multi-class classifier for the HLA problem

The original dataset contained 69006 samples, each one with 1883 features. After a reduction process, the dataset to be used contains 59229 samples and 527 features. There are a total of 231 different possible labels for a sample (all the identified groups).

5.2.4. House price prediction Dataset

The final dataset to be considered, is a prediction dataset. Donated by the CMU Statlib library [12], it includes information regarding the housing costs in suburbs in the Boston area, as collected by the 1990 census. Each feature contains 12 real and one binary features (13 in total) and a prediction of the price of the house. The information that each feature provides is indicated in table 5.3.

Name	*Description*
CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PT RATIO	pupil-teacher ratio by town
B	$1000(Bk - 0{,}63)^2$ where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000's

Table 5.3: Value indicated by each feature. Obtained from [12]

The dataset is extensively used as benchmark comparison between learning algorithms , where the registered generalization error goes from around 8 % [6] to around 3.64 % [25]. It contains 506 training instances.