Investigating the Structural Diversity within a Committee of Classifiers and their Generalization performance

Lesedi Melton Masisi

A dissertation submitted to the Faculty of Engineering and the Built Environment, University of the Witwatersrand, Johannesburg, in fulfillment of the requirements for the degree of Master of Science in Engineering

Johannesburg, 2008

Declaration

I declare that this dissertation is my own, unaided work, except where otherwise acknowledged. It is being submitted for the degree of Master of Science in Engineering in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other university.

Signed this_____day of_____20____

_____

Lesedi Melton Masisi

Abstract

This study investigates the measures of diversity within ensembles of classifiers. The use of neural networks is carried out in measuring ensemble diversity by the use of statistical and ecological methods and to some extent information theory. A new way of looking at ensemble diversity is proposed. This ensemble diversity is called ensemble structural diversity, for this study is concerned with the diversity within the structure of the individual classifiers forming an ensemble and not via the outcomes of the individual classifiers. Ensemble structural diversity was also induced within the ensemble by varying the structural parameters (learning parameters) of the artificial machines (classifiers). The importance or the use of these measures was judged by comparing the measure of structural diversity and the ensemble generalization performance. This was done so that comparisons can be made on the robustness of the idea of structural diversity and its relationship with ensemble generalization performance. It was found that diversity could be induced by having ensembles with different structural and implicit (e.g learning) parameters and that this diversity does influence the predictive ability of the ensemble. This was concurrent with literature even though within literature ensemble diversity was viewed from the output as opposed to the structure of the individual classifiers. As the structural diversity increased so did the generalization performance. However there was a point where structural diversity decreased the generalization performance of the ensemble, where from that point onwards as the structural diversity increased the generalization performance decreased. This makes sense because too much of diversity within the ensemble might mean no consensus is reached at all. The disadvantages of comparing structural diversity and the generalization performance (accuracy) of the ensemble are that: an ensemble can be structurally diverse even though all the classifiers within the ensemble approximate the same function which means in this case structural diversity is meaningless in terms of improving the accuracy of the ensemble. The use of ensemble structural diversity measures in developing efficient ensembles still remains to be explored. This study, however, has also shown that diversity can be measured from the structural parameters and moreover reducing the abstractness of diversity by being able to quantify structural diversity making it possible to map a relationship between structural diversity and accuracy. It was observed that structural diversity does improve the accuracy of the ensemble, however, within a limited region of structural diversity.

I would like to dedicate this work to my mother, Elva Modiegi Masisi, my father, Rathapelo Samuel Masisi, my sisters and brothers, my two other brother's in Christ, Suprice Hlatswayo and Oscar and the family of Jacob Kim: for their support, believing in me and inspiring me that the best way to triumph over difficulties in life is by learning and showing your faith through your deeds.

"We please God by what we do and not only by what we believe"
(James 2, 24).

"If you have good instruction it will help you to have even better sense and to live right. Education will help you to know even better"
(Proverb 9, 9).

Acknowledgements

The work described in this dissertation was carried out at the university of the Witwatersrand, in the school of electrical and information engineering during the year of 2008. I would like to thank the input of Dr Nelwamondo who most of the time told me that I needed to return to the drawing board. I would also like to thank the control group guys for their input and the informal discussions that we would have, for it was sometimes through these discussions that strategies of solving a problem were inspired. I would like to thank the input of Prof Marwala for his inspiration that also made it possible for me to do a Master's degree, he has enabled me to reach international exposure. Thank you for your extensive leadership and insight and your hunger for success. I would like to thank the following people for helping me with the proof reading of this dissertation: Rofhiwa Musehane, Linda Mthembu and Kefilwe Sethaba. It is well appreciated, I thank you. **Excelent!**

I would also like to extend my thanks to the National Research Foundation for funding this work.

Table of Contents

Table of Figures

List of Tables

Published Papers

The following papers were published from the work done in this research. One is published and the others are accepted for publication.

- L. Masisi, F.V. Nelwamondo and Tshilidzi Marwala, "The effect of structural diversity of an ensemble of classifiers on classification accuracy", *International Association of Science and Technology for Development: Modeling and Simulation* (IASTED), pp. 135-140, 2008.

- L. Masisi, F.V. Nelwamondo and T. Marwala, "The use of entropy to measure ensemble structural diversity", 6th *IEEE International Conference on Computational Cybernetics*, pp. 41-45, 2008.

- L. Masisi, F.V. Nelwamondo and T. Marwala, "Investigating Ensemble weight Distributions for Indicating Structural Diversity", *15th International Conference on Neural Information Processing of the Asia-Pacific Neural Network Assembly* (accepted), Springer, 2008.

Nomenclature

ANNs	Artificial Neural Networks (Classifiers)
D	Bias component
E	Variance component
ESD	Ensemble Structural Diversity
GAs	Genetic Algorithms
IDS	Identity Structure
MSE	Mean Square Error
MID	Militarised Interstate Disputes
MLP	Multi-Layered Perceptron

Chapter 1

1 Introduction

Diversity is a well understood social concept, within society people can be diverse in language, culture, race, etc, intuitively it could be seen within a group of different people but expressing it mathematically or being able to measure diversity poses a challenge, for it is a complex subject. Within machine learning context, diversity comes to play when one uses more than one artificial machine to do a prediction. One of the important key factors in machine learning is the generalization performance and diversity is regarded as one of the key concepts that contribute to the effectiveness and efficiency of a committee of artificial machines to have a good generalization performance. Diversity in general can be seen as a variation or differences within a group of people or artificial machines, it could be a combination of factors such as race and culture within a social context or a combination of different outputs of artificial machines. However, in this dissertation diversity will be seen as a variation or differences within the structural parameters of a committee of artificial machines.

This dissertation addresses the study of structural diversity, how one quantifies or measures structural diversity and what value will this unique analysis of structural diversity bring to current knowledge on the generalization performance of the ensemble. This intern will allow one the chance to extract knowledge from structural diversity in relation to the generalization performance of the ensemble. However in order to capture the relationship between structural diversity and the generalization performance, structural diversity needs to be a measurable quantity. This would add value, for it will make it possible for proper rules in constructing committees to be made. Such rules would enable one to have committees that can have a better generalization performance.

Given that diversity is necessary within a committee of ANNs for better performance [1-3]. The need for an induction of diversity within a committee of ANN can also be justified by the fact that, without diversity then all the ANN composed within a committee will have the same

decision boundary, meaning even if one ANN was used the same performance as that of the committee would be observed. The use of a committee of ANN is chosen so that different structural diversity topologies can be created which would then lead to the goal of this dissertation. There have been a number of measures of ensemble diversity that have been developed within the machine learning research community; such measures will be discussed in chapter 2.

The measures that have been developed focus on one stage called the output stage of the artificial machines in measuring diversity. This means for example if two artificial machines agree on an outcome then they are considered not diverse but if they had disagreed then they would be considered diverse. This view of ensemble diversity has brought about a number of mathematical formulations that have been able to quantify diversity as seen from the outcomes (the output stage). In light of these facts, this dissertation therefore introduces a new way of viewing and measuring ensemble diversity. Instead of measuring ensemble diversity from the output stage (outcomes) of the artificial machines, the learning parameters (structural parameters) of the artificial machines are used for quantifying ensemble diversity. This is inspired by the fact that there is no formal definition of diversity and the fact that this subject is still left as an open topic, one of Kuncheva's and Whitaker's conclusion was that: *"…the general motivation for designing diverse classifiers is correct but the problem of measuring this diversity and so using it effectively for building better classifier teams is still to be solved"* [4]. Such machines have different learning parameters which will be continuously regarded as structural parameters for the rest of this study. Having the knowledge of the diversity induced and how that diversity influences the accuracy of the ensemble will be of utmost importance, for one could use this as a way of constructing efficient committees of artificial machines. Hence this dissertation is concerned with the measure of the structural diversity of a committee of ANNs (MLP) and further to indicate that it can be used to improve the generalization performance of the committee. Section 1.1, titled ensemble systems, discusses the advantages of using a committee of classifiers, requirements for effective and efficient operation of a committee and a background on research done on measures of ensemble diversity and section 1.2 titled ensemble diversity and generalization performance, deals with the methods taken for measuring ensemble structural diversity and section 1.3 titled, data exploration, is concerned with exploring the complexity of

the data set used and section 1.4 titled structural diversity measures, is concerned with the representation of the classifiers and the structural diversity measures conducted and section 1.5 summarizes the outline of the dissertation.

1.1 Ensemble Systems

Ensemble based systems either for regression or classification problems deals with combining the artificial machines as opposed to using only one artificial machine to do a prediction. Ensemble based systems are systems that have long before been established, perhaps one of the earliest work done in ensembles is by Darsarathy and Sheela's [5] 1979 paper. Hansen and Salamon [6] showed that the performance of a neural network can be improved by using an ensemble composed of similar type of the neural network. Gordon Brown [7] also showed that error function in regression problems of an ensemble of artificial machines reduces the variance component of the decomposed error function, resulting in a reduced prediction error. It is evident that a committee of artificial machines could have a greater advantage over the use of only one artificial machine in terms of the generalization performance [8, 9,10, 1] and further studies are summarized in Dietterich [11].

Ensemble systems have appeared in literature with various names such as, multiple classifiers [12], combination of multiple classifiers [13], dynamic classifier selection, classifier fusion, [14-16], committees of neural networks and consensus aggregation among many others. These applications differ from each other depending on the procedure that generated the classifiers and the procedure for combining the individual classifiers. In general there are two approaches forming ensembles, classifier selection and classifier fusion. In the former each classifier is trained to be the best in some certain local area of the entire feature space. However in classifier fusion all the classifiers are trained over the entire feature space and then combined at the end.
In this dissertation both the above mentioned combination methods are explored. There are diverse reasons for using the ensemble based systems, one is that it is generally better to use an ensemble of un-optimized classifier parameters rather than optimizing a particular classifier for good generalization performance. Hence the use of ensemble system is to lessen the load on optimization techniques for constructing strong classifiers.

The aggregation or fusion method is done by combining all the trained classifiers to form a stronger classifier similar to a bagging or boosting based approaches. Researchers have shown large interest into ensemble of artificial machines, the training methods and the aggregation schemes [9, 14, 17]. Studies have shown that there is a correlation between the generalization performance of the classifiers with the way the outcomes of the classifiers are combined [14]. The aggregation schemes have created large interest when compared to ensemble generalization performance (accuracy) of the ensemble [18, 19]. An aggregation scheme is a method of combining a number of classifiers so as to produce a unified result, meaning that out of many classifiers, a strong classifier is made. Strong in this case would imply a classifier that can generalize better than one classifier. There are a number of combination schemes such as, majority vote, averaging, combination of posterior probabilities, etc. The other challenge in working with ensemble systems is the combination method scheme, for they also have an effect in the generalization performance of the ensemble [14, 17, 20, 21, 13, 22]. In this dissertation the majority vote combination scheme is used due to its wide use and is believed to have some form of neutrality. A study on the use of other combination schemes is beyond the scope of this dissertation. This will then allow more emphasis on ensemble diversity and the generalization performance and would reduce the complexity of this study.

One of the key concept within ensemble systems is the diversity of the classifiers within the ensemble, studies have shown that artificial machines within the ensemble need to be diverse for the ensemble to be efficient (perform better than an individual machine) [3, 20, 1, 19]. This is normally done by combining the individual classifiers with different outcome errors. This leads to the creation of an ensemble composed of classifiers which have different decision boundaries [1]. However, for this dissertation classifiers with different structural parameters will be combined and such an ensemble will be considered diverse. Ensemble systems have led to research into measures of ensemble diversity with the aim of understanding ensemble diversity so as to be able to build effective ensembles as mentioned. This is assessed by comparing ensemble diversity measures and generalization performance of the ensemble, which could result in a robust and generic function that relates ensemble diversity and generalization performance (accuracy). This would then assist greatly in the construction of efficient and effective ensembles

and hence has inspired the investigation of structural diversity with respect to the generalization performance as opposed to diversity as seen from the outputs of the individual classifiers.

In this research, many ensemble diversity measures have focused mainly on measuring ensemble diversity from the outcomes [22-24] (By the outcomes it means the final output made by the individual classifiers). These measures include, the Yule's Q-static for two classifiers, correlation coefficient (ρ), Kohavi-Wolpert variance (kw), Entropy measure (Ent), measure of difficulty (θ) and Coincident Failure Diversity (CFD) [3]. Researchers have tried to correlate these ensemble diversity measures with ensemble generalization performance and not all the measures correlated well with the generalization performance [1, 18, 19].

This dissertation proposes a new way of looking at ensemble diversity, whereby instead of looking at the classifiers at the outcome for measuring ensemble diversity one looks at the structural composition of the classifiers in quantifying ensemble diversity. This is inspired by the fact that in current literature in ensemble diversity measures,

1. There has not been one that is officially accepted for use in developing ensemble learning algorithms.
2. Ensemble diversity has no formal definition, which therefore leaves room for other exploration in ensemble diversity measures

Ensemble diversity in this context is defined as:

the ensemble that is composed of classifiers with different structural parameters as opposed to the one with the same structure.

This of course leads to the consideration of the induction of diversity and the way the classifiers are trained and how the generalization performance is measured. The measures used will then be weighted by comparing them to the generalization performance of the ensemble. This is called the validation process. The validation process will then validate or bring about confidence on the ensemble diversity measure used. It is this validation process, which forms the link between diversity measures and the generalization performance that has geared up challenges within the diversity measures research community [4, 25]. The data used to showcase the work done in this dissertation is the interstate conflict data. A similar work of measuring ensemble diversity and

then comparing ensemble diversity with the generalization performance of the ensemble was done, however, using the HIV dataset [26], the method used is similar to the one conducted in chapter 4. Section 1.3 will showcase the complexity of the data which hence necessitates the use of artificial neural networks. The following section looks at the relationship between ensemble diversity and the generalization performance.

1.2 Ensemble Diversity and Generalization Performance

One of the major challenges in this work is the representation of the artificial machines so as to implement the diversity measure, since the representation influences the mathematics that has to be implemented in quantifying structural diversity. Ensemble diversity can be induced in a number of ways such as applying different training algorithms (Boosting, Bagging, etc)[1, 27] , and by a committee composed of classifiers of different structural parameters. The relationship between ensemble structural diversity measures might not relate well with the generalization performance of the ensemble, due to:

1. An ensemble might be composed of classifiers of different complexities, which might all approximate the same output function.
2. Classifiers have a number of inherent parameters such that if all are not observed the structural diversity measure might just be measuring an index of structural diversity, which might not relate well with generalization performance.

This means that to reduce the risk of this challenge, the structural parameters of concern must be very influential in the generalization performance of the classifiers. Hence little work will be done on the consideration of all the structural parameters and then more work done on the hidden nodes as they were noted to be highly influential in the generalization performance of the ensemble [28]. Hence this study aims to develop a measure of structural diversity induced on the ensemble and then evaluate the induced diversity with the generalization performance of the ensemble. In this way we would be measuring the ensemble diversity that has been induced and then question on how this induced diversity affects the generalization performance of the ensemble. This ultimately might lead one to the knowledge of how structurally diverse of an

ensemble should be for better generalization and further this would have quantified the induced structural diversity of the ensemble. However due to the use of one dataset, this dissertation will mainly emphasize the existence of structural diversity within the ensembles by having a form of a measure and then extract knowledge to show how this diversity affect the generalization performance (accuracy) of the ensemble. This will result in a relationship between structural diversity and generalization performance (accuracy) being captured. Hence this dissertation does not necessarily aim to reach a highest accuracy. The following section explores the data used in this work.

1.3 Data Exploration

The autocorrelation of the (MID) data has been performed to give a better understanding of the data. This tests if the features of the dataset are sensitive to changes due to other features. This correlation is expressed by the use of |r|, within the range of between -1 and 1. Information is extracted from the absolute value of r. That means for, $|r| \rightarrow 1$, the more the variables observed correlate and if, $|r| \rightarrow 0$, then there is no correlation. The (+) and the (-) indicate a positive and a negative correlation respectively. The equation that defines the covariance between two variables can be defined as [29]:

$$cov(x, y) = \frac{1}{N}\sum_{i \in I}(x_i - \bar{x_i})\,(y_i - \bar{y_i}) \tag{1.1}$$

Where, $\bar{y}$ and $\bar{x}$ are the average values of y and x, respectively and N is the total number of samples. The correlation coefficient is the scale-variant of the covariance coefficient. By normalizing the covariance coefficient by the standard deviations, the correlation coefficient can be found as:

$$r(x, y) = \frac{cov(x,y)}{(s(x)s(y))} \tag{1.2}$$

where $s(x)$ and $s(y)$ are the variances of the variables x and y, respectively. The correlation coefficients of all the variables with each other can be seen from table 1.1.

Table 1.1: The correlation coefficients of all the variables with respect to each other

	U1	***U2***	***U3***	***U4***	***U5***	***U6***	***U7***	***U8***
U1	1	0.18442	-0.053826	-0.019904	0.034152	0.27659	0.072781	-0.18271
U2	0.18442	1	0.2947	-0.3304	-0.20803	0.10994	-0.32909	0.01862
U3	-0.053826	0.2947	1	-0.68849	-0.55041	0.15076	-0.7846	0.46415
U4	-0.019904	-0.3304	-0.68849	1	0.50322	-0.20625	0.70804	-0.38495
U5	0.034152	-0.20803	-0.55041	0.50322	1	-0.20091	0.55319	-0.37879
U6	0.27659	0.10994	0.15076	-0.20625	-0.20091	1	-0.094343	-0.04316
U7	0.072781	-0.32909	-0.7846	0.70804	0.55319	-0.094343	1	-0.36835
U8	-0.18271	0.01862	0.46415	-0.38495	-0.37879	-0.04316	-0.36835	1

The last column represents the outcome from a dispute. These results show that the data is not linearly correlated even the correlation between the individual inputs and outputs are very small. Hence the use of Computational Intelligence (CI) methods will be considered in this study, since they are capable of capturing the dynamics of the non-linear variables, a background on the tools used can seen from chapter 2.

1.4 Structural Diversity Measures

Diversity in this context is defined as having an ensemble that is composed of ANN that have different structural parameters, such structural parameters include, the number of hidden nodes, learning rates, activation functions. Three methods of indicating structural diversity have been considered.

- *Binary representation*

The first method looks at giving the artificial machines a unique identity (representation), the machines are given a certain binary code. This is inspired from the fact that each individual has a unique gene structure. Due to this representation a statistical measure was made that quantified structural diversity. The Kohavi-Wolpert variance uses such a representation to quantify diversity.

- *Species representation*

The second method adopted the ecological concepts by viewing the classifiers as species in quantifying structural diversity.

- *Parameter distribution representation*

This method looks at the distributions of the classifiers weights and biases as an indication of structural diversity, this was because a comparison was done between ensembles that were diverse due to a training algorithm and an ensemble that had different structural parameters. A brief introduction of these measures is given in the following sections. A detailed explanation and implementation of the mentioned measures will be done in chapter 3, 4 and 5 respectively.

1.4.1 Ecological Measures

The measures employed within population of species in ecology are to calculate the frequency or proportion of different species with a certain area. In this context the ANN with different structural parameter will be treated as different species. In ecology, the use of entropy to count species is applied. Hence for this study, entropy measure will be used to quantify structural diversity of the ensemble. Three entropy measures will be employed: the Shannon, Simpson's and the Berger Parker entropy measures. The hypothesis behind the success of this measure is that, structural diversity is induced by having an ensemble with different structural parameters (different species), hence by using entropy measures, one will be measuring the structural diversity induced, due to the way it was induced.

1.4.2 Kohavi-Wolpert Variance

It has been noted that the structural diversity measure suffer from the representation of the individual classifiers in such a way that it would be measured. The Kohavi-Wolpert Variance is a statistical measure that has been applied on the outcomes to measure ensemble diversity [1]. However in this context, it is interpreted differently. It is used to measure the structural diversity of the ensemble. It is a variance measure, as the variance increases so does the structural diversity of the ensemble. The classifiers are then given an identity representing their structure and this mimics the way gene string represent a structure or state of a cell.

1.4.3 Weights Distributions

In this work, the variances of the distributions of the individual classifier's weight vector (biases and the weights) are studied. The work done in this chapter is different from other chapters in the sense that diversity was induced by structure and also by a learning algorithm such as Bagging. Thus by considering the distribution of the weights and biases we can indicate the ensemble structural diversity. A comparison is then conducted between the non-diverse ensemble and the diverse ensemble in terms of their generalization performance.

1.5 Outline of Dissertation

This study aims to showcase a new approach to ensemble diversity measures. These measures are based on the structure of the classifiers and not the outcome (error diversity). This view has introduced the name ensemble structural diversity. However, to gain insight from these measures, their significance into the contribution of a search for measures of ensemble diversity so as to build efficient learning algorithms will be validated by relating the measures of structural diversity and generalization performance. The dissertation will, therefore, use an MLP and the interstate conflict data to demonstrate this concept. Diversity measure alone does not pose a problem, however, when compared to generalization performance; some measures of error diversity have produced poor results. Hence three approaches are employed in indication ensemble structural diversity. The experimental results conducted are given in a form of papers that are published and some that have been accepted [30-32]. A brief outline of the dissertation is provided below:

Chapter 2 provides the background on the modeling tools and the applications of ensemble methods, different types of error diversity measures used and the importance of the aggregation methods. A thorough theoretical background on ANN and the GA is given. A brief discussion on the advantage of error diversity measures over the structural diversity measures is also introduced.

Chapter 3 presents the use of ecological methods to quantify structural diversity. The classifiers are seen as species and the diversity index measures derived from entropy are implemented to quantify the structural diversity. The GA is used and the evaluation function is meant to optimize the classification accuracy. The diversity indices found are then compared with the generalization performance of the ensemble.

Chapter 4 explains the use of the Kohavi-Wolpert variance to quantify structural diversity. The GA is used, however, in this case to optimize the structural diversity values from a large ensemble. The GA searches for 21 classifiers that would give a certain structural diversity. In this section diversity is induced by having an ensemble with different structural parameters. The measure of structural diversity is implemented as mentioned and then compared with generalization performance of the ensemble.

Chapter 5 compares a structurally diverse and a non-diverse ensemble, by observing the variances of the distribution of the classifier weights. As opposed to chapter 3 and 4, ensemble diversity in this case is induced by the learning method. Hence the structural diversity is measured from observance of the distributions of the weights.

Chapter 6 summarizes the findings and provides some recommendations for future work.

Chapter 2

Background on Ensembles and Modelling Tools

2.1 Introduction

An ensemble in this context is a combination of classifiers. This chapter presents a literature review on the use of ensembles and their applications in other areas and the artificial intelligence tools that have been used to understand the behavior of ensembles with regard to the generalization performance. One key component in ensemble construction is the diversity within the classifiers. This resulted in a number of ways on measuring ensemble diversity, however, focused mainly on one area within the ensemble diversity research. The research done on ensemble diversity has been focused on the decision stage (at the outcomes). However, the outcomes give the final results. However, in this dissertation the aim is to measure ensemble diversity that resulted due to the structural composing of the committee and then extract knowledge on how ensemble structural diversity relates with the generalization performance. Artificial intelligence tools such as the Genetic Algorithms (GAs) will be used to gain such a relationship. The hypothesis is that an ensemble that consists of structurally diverse artificial machines would produce an ensemble that is composed of machines with different decision boundaries or an ensemble that has significant un-correlation among the classifiers forming the ensemble.

The remainder of the section is as follows, section 2.2 presents a literature review on ensemble diversity, and section 2.3 presents a background on neural networks and the effect of using ensembles and section 2.4 presents the background on the Genetic Algorithms (GA) which are used extensively in this dissertation to gain the relation between structural diversity and generalization performance.

2.2 Literature Review

Ensemble of artificial machines has widely been used, for it has been found that they perform better than the use of a single machine [9, 11, 13, 24, 33, 34]. This has led to the use of an ensemble of artificial machines in a variety of applications. For example ensembles have been used for: neural network learning [8], pattern classification [35], predicting HIV protease cleavage sites in proteins [36], decision making [1], data fusion [12], classification of prostate cancer [37] among others. Ensemble systems have had huge roles in machine learning applications and still continue to attract researchers [11]. It is therefore evident that the efficiency of ensemble systems on better generalization performance out performs that of the use of single machines. Researchers have developed better learning ensemble algorithms so as to optimize the generalization performance of ensembles [10, 21, 38, 39].

There are conditions to the effective operation of ensembles and one of the key components is that the individual classifiers be uncorrelated [1, 11] and this means that having an ensemble that is composed of artificial machines that have different decision boundaries. In other words the ensemble needs to be diverse. This has lead to a number of algorithms such as bagging, boosting, learn++ [3, 13, 39, 40] that induce ensemble diversity. There has been a number of developments in measures of ensemble diversity with the hope of using them in constructing more efficient ensemble learning algorithms among the ones already developed. This has resulted in a number of ensemble diversity measures [17, 20, 30, 41], however, these measures focus on the outcomes of the individual classifiers as mentioned. Such measures include: pair wise, disagreement and double fault, entropy, Kohavi-Wolpert variance and measure of difficulty among many [1, 20].

Kuncheva and Whitaker [4] have found that some ensemble diversity measures were highly correlated with the generalization performance of the ensemble [20]. They compared ten different measures of ensemble diversity and concluded that the idea of constructing diverse ensembles was correct, however, the measure of this diversity and then using it to construct efficient ensemble still remained to be explored and the fact that the concept of ensemble diversity is not clearly defined [4]. This means that there could be other ways of viewing and calculating ensemble diversity other than at the classifier outcomes.

It has been shown through studies that it is not only diversity that attributes to efficient ensembles but the way the artificial machines are combined. This has led to research on aggregation schemes [9, 13, 14, 42], where some of the aggregation schemes were found to be well correlated with particular diversity measures. However for this study only majority vote will be considered due to its wide use.

Instead of focusing at the outcomes, one could focus on the machine's structures for measuring ensemble diversity [30]. One of the base justifications for this work is that, studies have shown that ensemble generalization performance can be improved by having an ensemble that is composed of machines with different structural parameters. This study aims to measure ensemble structural diversity and then assess its predictive capabilities by looking at how the measure correlates with ensemble generalization performance. Computational machines particularly artificial neural networks and the GA optimization tool will be used.

2.3 Neural Networks (NNs)

The Artificial Neural Network (ANN), is a computational intelligence tool that can be used to capture input-output relationships. A feed forward neural network will be used in this dissertation also known as the Multilayer Perceptron (MLP). They mimic biological neural systems. In figure 2.1, each neuron received information from the previous neuron, each of those signals are then multiplied by a weight value that links the signal to destination node in that layer. The weighted inputs are summed and the passed through a limiting function resulting in a scaled output between a certain boundary. The output of a limiter is then shared to all other neurons in the next layer. This propagation of signals continues until the final output node.

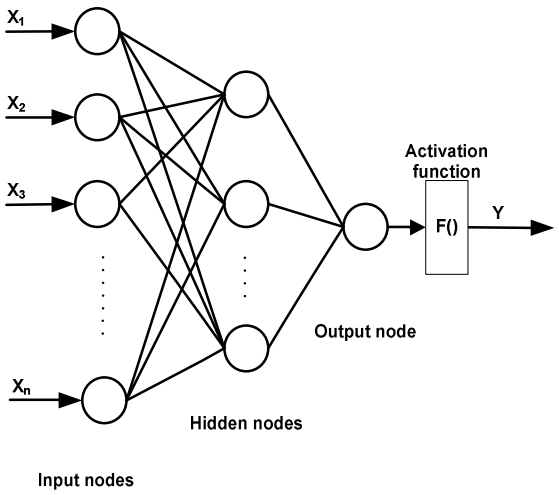

Figure 2.1: The MLP structure showing the inputs, the layers and the activation function [43].

The artificial NNs are currently being applied in almost every field in industry, bio-medical, financial institutes, risk analysis [44-47], condition monitoring which leads to the prediction of the life processes, optimization of business, conflict management, etc [43, 46, 48]. The use of a committee of artificial NNs has been used for improving the generalization performance of classifiers. ANNs are artificial machines that have the ability to map inputs to the outputs. The NNs are capable of deriving a function that defines a complex linear or non-linear behavior through their learning process. There are two types of learning paradigms, which are referred to as supervised or unsupervised. Supervised learning is when the NN is given the outcome of the input feature, during the training process while unsupervised learning is when it is not given the outcome of the input feature. For this study supervised learning was conducted.

Inputs are weighted and then summed into the limiting function before the weighted signals could be activated to the next layer of neurons as can be seen in figure 2.1. The weights are the most important parameters of the NNs for inducing intelligence within the NNs. Hence this study will extensively be concerned with inducing diversity and measuring ensemble structural diversity. A method of training the NNs that was conducted in this study was by means of Back-Propagation (BP). Hence the BP algorithm will allow the NN to adapt through a process of minimizing the error between the outputs of the NN and the correct data outcome.

However, the problem occurs when the problem domain is too complex such that it becomes even difficult to train one NN so as to produce a good generalizing network (accuracy). This might result in undesirable optimization techniques on the learning parameters of the ANN, given the fact that even when they are optimized sometimes one machine (ANN) is not capable of capturing the dynamics of larger problem domains [4]. This is one of the reasons ensemble based systems have been recommended to outperform the use of one NN regarding their efficiency and effectiveness. Ensemble based learning algorithms are similar in some way to the BP algorithm, only that BP operates at a low level dealing with the individual weights that constitutes the NN. The BP algorithm compares the output of the NN and that of the correct outcome for that instance, and then back propagates the error into the network and then small changes are made on the weights in each layer. The weight changes are done with the goal of minimizing the error between the output of the NN and that of the correct value for that input value as mentioned. This process is repeated until the overall error drops below a pre-defined threshold. The network is said to have learned some complex function that can map the inputs to the outputs. However the network does not learn the complex function exactly, it will have approximated the complex function to some degree. In ensemble based training paradigm, the training is such that individual NNs can be weighted depending on their performance [1]. This is the case for Adaboosting among many others. Equation 2.6 defines the process by which an input instance propagates to the output [43].

$$y_k = f_{outer}\left(\sum_{j=1}^{M} w_{kj}^{(2)}\, f_{inner}\left(\sum_{i=1}^{d} w_{ji}^{(1)} x_i + w_{jo}^{(1)}\right) + \; w_{ko}^{(2)}\right) \tag{2.6}$$

where f_{outer} and f_{inner} are the activation functions at the output layer and at the hidden layer respectively, M is the number of the hidden units, d is the number of input units, $w_{ji}^{(1)}$ and $w_{kj}^{(2)}$ are the weights in the first and second layer respectively, $w_{jo}^{(1)}$ and $w_{ko}^{(2)}$ are the biases of the first and second layer respectively when moving from input i to hidden unit j. By having an ensemble composed of classifiers that have different learning and limiting parameters (structural parameters) would imply a structurally diversity ensemble. The ensemble is considered

structurally diverse since it is composed of different classifiers with different structural parameters. This is where the concept of structural diversity stems from.

2.3.1 Variance Reduction

There is no formal definition of ensemble diversity, in this study it's been seen as the differences within the classifier's structural parameters. However, it appears that many researchers have looked at measures of diversity from the outcomes of the artificial machines as mentioned. Diversity, however, is acclaimed to improve the generalization performance. This was shown by Tumer and Ghosh [43, 49] by bombarding one of the measures used to quantify the performance of the artificial machines, the mean square error. It was decomposed into a bias and a variance component [50, 51]. See equation (2.7),

$$MSE = E + D \tag{2.7}$$

Where: E and D are the variance and bias components, respectively. They found that ensemble diversity reduces the variance component thereby reducing the overall error of the committee. Researchers have looked at the outcomes of the individual ANNs and if these individual outcomes where different then the ensemble was regarded as diverse. This dissertation, however, views diversity from the structural point of view and not the outcomes which means that one would be looking at the variations within the structural parameters of the ensemble. Hence an investigation on the relationship between structural diversity and the MSE is in question. Now it is well established that the number of hidden nodes plays a major role in the complexity of the artificial machines [43]. A measure that looks at the variations within the hidden nodes among the classifiers forming an ensemble as a form of expressing structural diversity of the ensemble form a big part of this study. Hence diversity will be induced with a goal of minimizing the variance component depicted in equation (2.7).

2.4 Genetic Algorithm

Genetic Algorithms are stochastic evolutionary search process that was invented by Holland (1975) [52]. They are evolutionary models that apply evolutionary biology. They make use of biological processes such as mutation, natural selection, reproduction and crossover [52, 53]. The GA finds the best individual (chromosome) by evaluating them via some cost function that relates the optimization problem with the GA [53]. This normally occurs through a randomly generated population of individuals and occurs in a number of generations. The stochastic nature of the GAs allows them to search through solutions to come up with a global maximum. See figure 2.2 for the sequence of the optimization process.

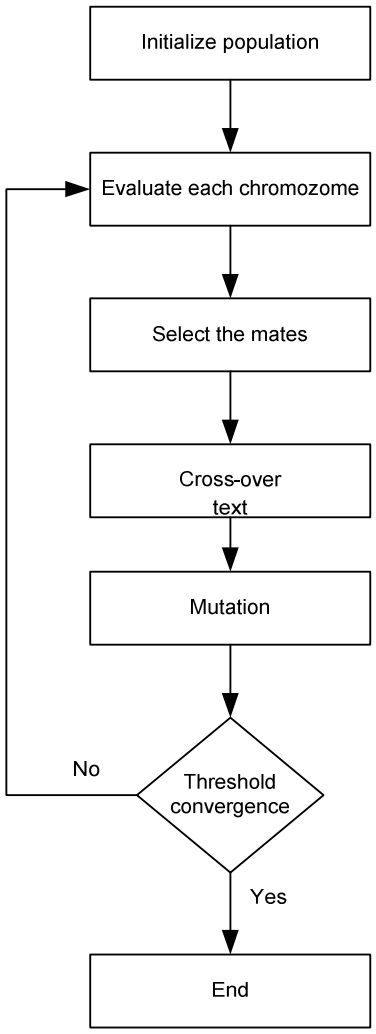

Figure 2.2: Flow sequence of the GA [54]

The algorithm will continue until the approximate fittest level has been found or if the maximum generation cycles are reached. In the implementation of the GA, a number of parameters need to be optimized. These parameters are: initial population size, termination threshold, the mutation rate and the crossover probability etc. In applying the GAs to the study of structural diversity, the evaluation function can be the structural diversity or the generalization performance. The idea is to relate structural diversity and generalization performance. If the evaluation function evaluates the structural diversity, then a measure for structural diversity will be conducted on a chosen number of classifiers making up an ensemble. The GAs would then search within a number of classifiers forming a group of them in order to attain the predefined structural diversity. Such a committee will then be related to the performance. If the evaluation function is the generalization performance, then the GAs will look for a defined number of classifiers that give that performance as a whole and then a diversity measure can be carried out at the later stage. This option also allows a relationship between the generalization performance and the structural diversity to be observed, see figure 2.3.

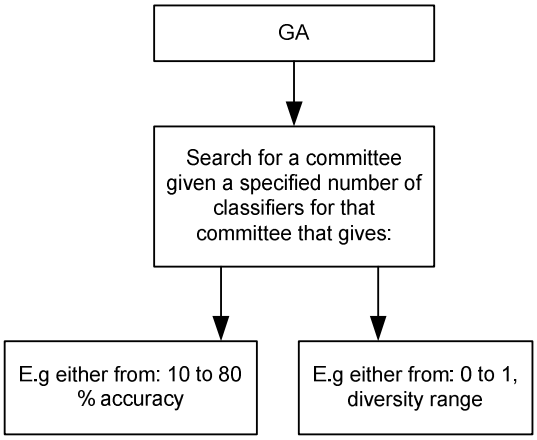

Figure 2.3: The use of GA in ensemble structural diversity

The committee will be given a specified number of classifiers and then the GA would find the optimal group or combination of the classifiers to either attain a certain ensemble generalization performance or ensemble structural diversity, as can be seen from figure 2.3.

Chapter 3

Ecological Methods to Measure Structural Diversity and Generalization Performance

3.1 Introduction

This chapter addresses the measure of structural diversity of an ensemble by using ecological measures. Diversity is induced by varying the structural parameters of the classifiers [30]. The three parameters of interest include the activation function, number of hidden nodes and the learning rate. This study aims to find a suitable measure of structural diversity by using methods adopted in ecology. The ecological measures are, therefore, aimed at bringing more knowledge to how ensemble structural diversity relates with the ensemble accuracy by quantifying structural diversity in terms of diversity indices. The ecological methods measure the index of diversity to quantify species diversity. These diversity indices are derived from Renyi entropy stemming from information theory hence these diversity indices for the rest of this chapter will be referred to as entropy measures, the derivation can be seen at section 3.3.

The classifiers will be treated as species in this chapter and hence quantifying ensemble structural diversity in terms of diversity indices. For example if there are three different species, two of the same kind and one of another kind, then that would be taken as three groups of MLP's of different structural parameters. However, this study will only focus on three measures of diversity indices, Shannon, Simpson and Berger Parker to quantify structural diversity of the classifiers. Shannon in this chapter will be viewed in two ways, firstly it will be used to measure uncertainty as adopted from information theory [55] and then secondly as a measure of diversity index as adopted in ecology for species counts [56]. The interest on uncertainty is so that ensemble structural diversity can be viewed in the context of uncertainty.

The relationship between the classification accuracy and the entropy measures or diversity index measures is attained by the use of Genetic algorithms using accuracy as the cost function [30]. This chapter includes a section on the background, Species and the Identity Structure (IDS), Renyi entropy, Shannon diversity index, Simpson diversity Index, Berger Parker index, neural network parameters, Genetic Algorithms (GA), the model, the data used, results and discussion and then lastly the conclusion.

3.2 Species and the Identity Structure

The ensemble of classifiers was treated as species and this was made possible by the ensemble being composed of different structural parameters. The different structural parameters that were varied were the: activation function, number of hidden nodes and the learning rates. Each classifier had a unique identity due to the different structural parameters used for each classifiers and this identity was called the Identity Structure (IDS), see chapter 4 for the extended explanation of the IDS.

$$IDS = \begin{bmatrix} Activation\ function \\ Number\ of\ hidden\ nodes \\ Learning\ rate \end{bmatrix}$$

Five learning rates were considered and three activation functions similar to [30]. The number of hidden nodes was between 7 and 21. It was made larger than the attributes (inputs) so as to have classifiers that could generalize well and then less than 21 so as to reduce the computational costs. The learning rates considered were: 0.01, 0.02, 0.03, 0.04, 0.05 and the activation functions were: The sigmoid, linear, and the logistic. This chapter is a continuation of [30], whereby the identity was converted into a binary string. In this chapter there was no need to convert the identity into a binary string since the entropy measure only looks at the machines which are different. The individual classifiers forming the ensemble were given different numbers as according to their identity. Defining an identity for each machine is necessary so as to have a unique identity of the classifiers within the ensemble. This will in turn enable the use of the uncertainty measure on the ensemble, for the IDS can be treated as a symbol representation representing a particular classifier due to its unique structure.

3.3 Renyi Entropy

Renyi entrony [57] is composed of the three measures of diversity mentioned in this chapter as follows.

$$H_{\propto} = \frac{ln(\sum P_i^{\propto})}{1-\propto} \tag{4.1}$$

where P is the proportion of an item i.

The diversity measures can be found by, Shannon ($\propto \rightarrow 1$), Simpson's ($\propto \rightarrow 2$) and the Berger Parker ($\propto \rightarrow \infty$). This means that as alpha ($\propto$) approximates the indicated values so does a certain measure get approximated (e.g Simpson's).

3.3.1 Shannon Entropy

Shannon Entropy [55] in information theory is perceived as the measure of uncertainty. If the states of any process cause the process after 10 iterations to give a series of ones, then one would be certain of the next preceding information [55]. However, if the states are diverse then we become uncertain of the outcome. Having an ensemble of classifiers which are all the same, would imply that if one of them were to classify a certain instance of input data, then with high probability all of them would classify the same object alike. However, the more diverse the ensemble become the more uncertain one is of the overall decision of the ensemble. This analogy was used to relate diversity and uncertainty in this chapter. In information theory, the uncertainty is seen as bits per symbol [55]. The uncertainty can be partially explained from the following equation, by using logs by using base 2 which means the unit of uncertainty is in bits otherwise with base 10 it is digits [55]. Classifiers with different structural parameters were treated as different symbols. If the classifiers had the same structure then they would be taken as the same symbol.

$$u_i = -\log\,(p_i) \tag{4.2}$$

$p = 1/M$ is the probability that any symbol appears [55] (with M being the number of symbols, M is taken as the number of classifiers). By analogy, when taking the classifiers as the symbols, p is the probability of choosing any classifier within the ensemble and u is the uncertainty. This

could be possible since the classifiers were represented uniquely by their IDS (each classifier had a unique identit). From this analogy it means that if p tends to 0, then it is highly unlikely that the ith classifier will appear meaning that equation (4.2) tends to infinity hence high uncertainty in that regard [55]. Likewise, if P tends to 1 then there are high chances that the ith classifier will appear resulting in a reduced uncertainty when assuming normalization from 0 to 1. Shannon's general formula for uncertainty, see Equation (4.3), exists when $\propto$ (see section 3.3) tends to 1 [55].

$$H_1 = -\sum_{i=1}^{M} P_i log(P_i) \tag{4.3}$$

The maximum of Equation (4.3) occurs when the structural diversities of the classifiers are equally likely (technically this would imply equally likely symbols). This means when P_i = 1/M for all classifiers within the ensemble, substituting this into Equation (4.3) will result in, *log (M),* which is perceived as species richness in ecology [56]. For this study the Shannon diversity index was normalized between 0 and 1 by dividing Equation (4.3) by log (M) the maximum possible diversity index or entropy. A 1 will imply the largest structural diversity, which when viewed from the ecological perspective would imply a high diversity index. A 0 result would mean no structural diversity.

3.3.2 Simpson's Diversity Index

When taking α to 2, the Renyi entropy approximates to [57]:

$$H_2 = -log\left(\sum_{i=1}^{n} P_i^2\right) \tag{4.4}$$

It is the probability of any two individuals drawn at random from a large ecosystem belonging to different species [58]. The inverse of this expression is taken as the biodiversity index, which means that H_2 increases with the evenness of the distribution which is the diversity index in this case. A 1 will represent more diversity and zero no diversity. The normalization was done by removing the *log* and then by using, 1 - H_2 so that as the evenness increases so does the diversity index.

3.3.3 Berger Parker Index

The Renyi entropy approximates to equation (4.5), when taking α to infinity [58]:

$$H_{\infty} = \frac{1}{p_i} \tag{4.5}$$

Where $p_i$is the probability of choosing a certain classifier with a diverse ensemble. H_{∞} gives the equivalent number of equally abundant species with the same relative abundance as the most abundant species in the system [58]. The Berger Parker index only considers the relative dominance of the most popular species, ignoring all the other species. The Berger Parker index was normalized between 0 and 1 by dividing H_{∞} by 21 the total number of the classifiers within the ensemble. A 0 implies no diversity and a 1 highly diverse, this is when the ensemble is composed of classifiers which have different structural parameters, that means no classifiers is the same with any other classifiers (no repeats).

3.4 GA

In this study, the evaluation function is the ensemble classification accuracy, the GA searches for a group of 21 classifiers that would minimize the cost function. That means an ensemble that will produce the targeted accuracy. The GA searches through already trained 120 classifiers evolving the artificial machines (classifiers) to attain the targeted accuracy. The evaluation function is composed of two variables, the ensemble accuracy and the targeted accuracy T_{acc}. Equation 4.7 is the evaluation function. The ensemble was chosen to have 21 classifiers, the number was made odd so that there would not be a tie during voting and 21 was chosen arbitrarily.

$$f_{GA} = -(Acc - T_{acc})^2 \tag{4.7}$$

Where, f_{GA} is the evaluation/objective function, Acc is the accuracy of the 21 classifiers and T_{acc} is the targeted accuracies.

The GA tries to optimize the valuation function. Equation (4.7) will reach its maximum when the accuracy of the ensemble is equal to the targeted accuracy. GA was then optimized by first

searching the target values which the GA could attain. These were then the targeted accuracy values for the cost function for the next run. This was done so as to reduce the computational cost since the search space will be minimized. In other words for an example, the algorithm will search for a list of accuracies, 50 %, 55%, 60%, 65 %, etc from a combination of 21 classifiers and for each targeted ensemble classification accuracy an ensemble diversity measure was undertaken. For each classification accuracy attained there would be a quantified diversity measured. This would then make it possible for a relationship between ensemble diversity and classification to be captured.

3.5 The Model

The model describes the use of GA tool in selecting 21 out of 120 classifiers so as to provide knowledge of how the accuracy of the ensemble relates with the uncertainty of the ensemble. Figure 4.1 illustrates the use of 120 classifiers in attaining an optimal ensemble for classification. A method of voting is used to aggregate the individual decisions of the classifiers within the ensemble.

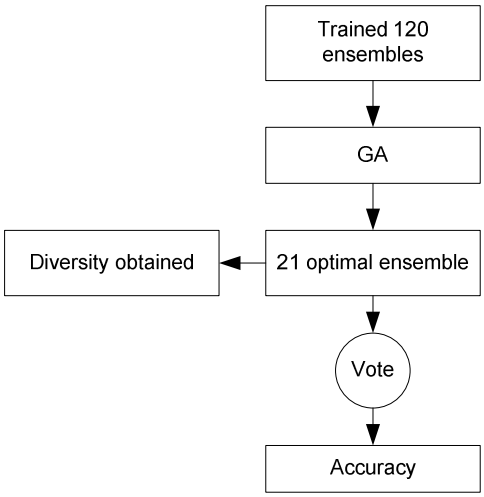

Figure 4.1: The method used to optimize the 21 classifiers of the 120 classifiers

3.6 Ensemble Generalization

The classification accuracy of the ensemble was attained by using a method of voting to aggregate the individual decision of the classifiers. For every classification done on the data sample, the number of correct classification was counted. Equation (4.8) is used for calculating the classification accuracy of the ensemble.

$$Acc = \frac{n}{N} \tag{4.8}$$

where n and N is the number of the correctly classified samples and the total number of the data samples to be classified, respectively

Classifier outputs from greater or equal to 0.5 where rounded to 1 and anything less than 0.5 was rounded to 0. This was because the outputs from the neural networks were taken as probabilities.

3.7 The Data

The data was normalized to fall between 0 and 1, to have equal weight of all the input features by using equation (4.9).

$$X_{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{4.9}$$

where x_{min} and x_{max} are the minimum and maximum values of the attributes of the data samples observed, respectively.

3.8 Results and Discussion

The entropy measures were calculated on the ensemble of 21 classifiers. These measures are quantified as the diversity indices of the ensembles. These are the results of 11 ensembles (each ensemble contains 21 classifiers) as were selected by the GA.

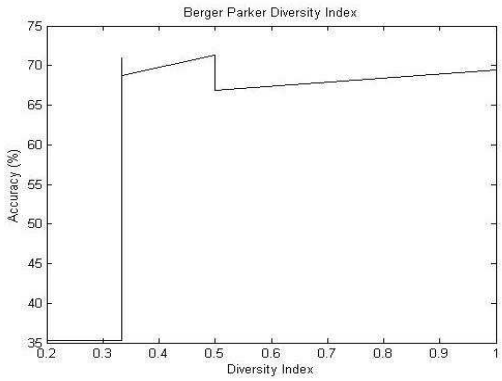

Figure 4.2: The Berger Parker index of diversity Vs ensemble classification accuracy

The Berger Parker measure is in agreement with the fact that one needs a diverse committee as opposed to the same type of classifiers within the committee. This can be seen as shown in the diversity indices of less than 0.34 whereby the classification accuracy is less than 37 percent, see Figure 4.2. As the diversity index increases so does the ensemble structural diversity become more evenly distributed.

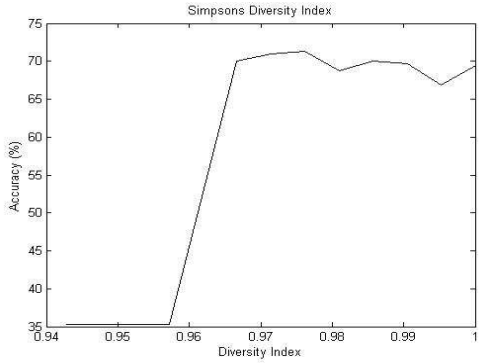

Figure 4.3: The Simpson's diversity index Vs ensemble classification accuracy

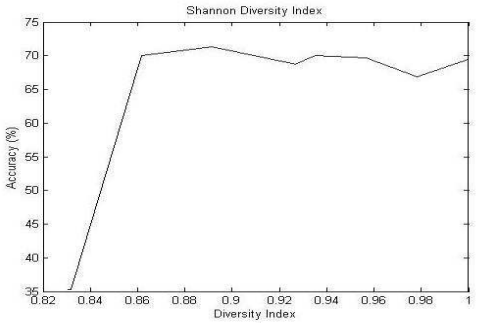

Figure 4.4: The Shannon diversity index Vs ensemble classification accuracy

All the diversity measures are concurrent with the literature regarding the diversity of the committee of classifiers as presented in [3]. The Shannon diversity index indicates that at very low diversity index, the generalization of the ensemble is poor, however, as the diversity increases so does the accuracy. There seems to be a high correlation between the Shannon and the Simpson's diversity indices in relation to the classification accuracy. The results from the Simpson's measure shows to be more sensitive to high diversity indices as seen in Figure 4.3. When considering the Shannon as the uncertainty measure, it can be seen that as the ensemble becomes more uncertain, its generalization ability increases. The Shannon diversity index and the Simpson's diversity indices have a decreasing accuracy after reaching a peak accuracy level, see Figure 4.3 and 4.4. This indicates that evenness on the classifiers needs to be limited for good ensembles. The use of accuracy as a function of Berger Parker diversity measure did not show to be a good function of Berger Parker measure of structural diversity of the ensemble. This can be seen in Figure 4.2 for the graph does not show functional properties.

3.9 Conclusion

This chapter presented the use of methods inspired from ecology which were derived from entropy to quantify structural diversity. These diversity measures were then compared to ensemble classification accuracy. Three measures of diversity indices were compared and it was observed that the ensembles accuracy improved as the structural diversity of the classifiers increased. The other interesting observation was that of the Shannon diversity index when interpreted as the uncertainty measure from the information theory. As the uncertainty of the ensemble increase so did the classification of the ensemble. This implies that having more information (as defined by Shannon's measure) of the ensemble might result in poor generalization ability of the ensemble, hypothetically. The method used to compute the results was computationally expensive due to the use of GA. Diversity indices from 0.82 to 1 were captured and the Berger Parker measure was not observed to be a good indicator of diversity when compared to the generalization performance (accuracy), for it did not produce functional properties (i.e there were still a number of diversity measured within the same classification

accuracy of 70 %). This chapter has also shown that Entropy based methods can be used to better understand the ensemble diversity in particular ensemble structural diversity.

Chapter 4

Ensemble Structural Diversity Measure and Generalization performance

4.1 Introduction

This study focuses on what was proposed by Sharkey [28], that diversity can be induced by having an ensemble of classifiers with different architectures. A method of identifying uniquely the individual classifiers is critical for quantitatively imposing a defined structure for each classifier. Since the structure of the individual classifiers is the focus of the study, the same data will be used to train the ensemble as opposed to bagging and boosting methods [3] in sampling the data for training, this is done so that only the architecture parameters of the classifiers would induce diversity. The Kohavi- Wolpert variance (KWR) [1] method was used to measure the structural diversity of the ensemble. The GA was used to develop a relationship between structural diversity and ensemble classification accuracy.

There are a number of aggregation schemes such as minimum, maximum, product, average, simple majority, weighted majority, Naïve Bayes and decision templates to name a few, see [17] and [35]. However, for this study the majority vote scheme was used to aggregate the individual classifiers for a final solution, due to its neutrality and wide use. An ensemble of 60 classifiers was created randomly and then trained. The GA tool was used to select a suitable group of 9 classifiers that would produce a targeted diversity measure as defined by the (KWR).

One of the other goals of this study is to map the relationship between structural diversity and accuracy. This will lead to knowledge of using the correct grouping of the classifiers for a desired accuracy or generalization ability. This chapter includes a section on the Identity structure (IDS), Kohavi-Wolpert Variance Method (KWR), data preprocessing, Genetical

Algorithms (GA), the model, Implementation, Results, Future work and then lastly the conclusion and discussion.

4.2 Identity Structure (IDS)

The Identity Structure (IDS) is derived from taking into account the parameters that make up an Artificial Neural Network (ANN). These parameters include the activation functions, number of hidden nodes and the learning rate. Other types of the ANNs can also be used to achieve the IDS. A number of artificial machines can therefore be used for a hybrid ensemble. Depending on how many different machine types one uses, the first, second, third, etc indexes of the IDS could stand for a Multi Layered Perceptron (MLP), Radial basis function (RBF), Bayesian radial basis function (BRBF), etc. That means if the index one of the IDS is a one, then the artificial machine would be an MLP for an example.

The IDS therefore demands a form of commonality between the artificial machines (activation function, hidden nodes, etc), because it represents the blue print for the individual classifiers that make up the ensemble. However for this study only one artificial machine was considered, an MLP. The IDS can be viewed as:

$$IDS = \begin{bmatrix} Machine\ type \\ Number\ of\ hidden\ nodes \\ Activation\ function \\ Learning\ rate \end{bmatrix}$$

The number of hidden nodes is set not to exceed 30, as shown by the five bold bits on the ID below. This conversion makes the ID less complex; however the number of bits can be increased.

ID = [1 **0 0 0 0 0** 0 1 0 1 1 0]

Each of the parameter of the IDS will have to be evaluated for measuring differences between the identities of the classifiers. The methods used to measure diversity are as follows: the Yule's Q-static for two classifiers, correlation coefficient (ρ), Kohavi-Wolpert variance (kw), Entropy

measure (Ent), measure of difficulty (θ) and Coincident Failure Diversity (CFD) [3]. These methods are mainly applied at the outcome of the classifiers and not at the building blocks of the classifiers (structure) [3]. However the Kohavi-Wolpert variance (*KWR*) method can be applied to measure the structural diversity, which was derived from the variance formulation [51].

4.3 Kohavi-Wolpert Variance

This method is applied in measuring the variance of the outputs of the classifiers in the ensemble and it falls under the family of non-pair-wise measures [3]. However, for this study it will be used to measure the variance of the different identities of the classifiers. That means for this study:

$$l(V_j) = \sum_{i=1}^{L} D_{i,j} \tag{3.1}$$

Where V_j, is a vector of the classifiers, L is the total number of classifiers belonging to either the RBF, MLP, etc family. That means if j = 1, then the ensemble is evaluated on the number of RBF machines present in the ensemble. V_j can be viewed as, $V_j = \left[ID_{1,j}^{T}, \ldots, ID_{L,j}^{T}\right]$. Equation (3.2) defines the overall variance calculation of the ensemble.

$$kw_r = \frac{1}{NL^2} \sum_{j=1}^{N} l(V_j)\left(\mathrm{L} - l(V_j)\right) \tag{3.2}$$

j = 1,…,N, where N is the number of the identity parameters (classifier type, complexity, and activation functions). This will result in the variance of the ensemble.

4.4 Data Preprocessing

A data sample of 1006 was used for training, 317 samples for validation and 552 for testing. The total data used was therefore 1875. This data has seven feature inputs as mentioned, however the data was normalized to have equal importance of all the features. The data was normalized to range between 0 and 1, by using equation (3.4):

$$X_{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{3.4}$$

Where x_{min} and x_{max} are the minimum and maximum values of the features of the data samples observed, respectively.

4.5 Genetic Algorithm

The genetic algorithm makes use of methods inspired from evolutionary biology such as crossover, reproduction and natural selection processes [52] and such methods can be used to evaluate certain functions. In this study the evaluation function is the diversity measure, the GA tries to meet certain diversity (KWR) among the ensemble of 9 classifiers. The chromosomes are the indexes for the vector that contains the classifiers. The GA will then evolve the classifiers for a specified diversity value. The GA faced difficulties in attaining the specified diversity. This was because the diversity measure specified could not be attained from the current ensemble of 60 classifiers which were arbitrarily chosen. To prevent this problem from occurring one would need to:

- Build the ensemble of 60 classifiers with known KWR values for any possible combination of the 9 ensembles.
- Initially run the GA for any KWR values and then use the set of KWR values that the GA can approximate.

The second option seems to be much feasible than the first option because on the first option it would mean that there would be no need for the GA. The first option further implies that the GA would be synchronized with the KWR measure. This was because it was also observed that some KWR that were obtained via trial and error were not able to be attained empirically by using equation (3.2), meaning that the GA could not converge. The second option was then used for implementing the GA. The GA was empirically optimized for an initial population of 20 chromosomes, 28 Generations with a crossover rate of 0.08. The optimization was done by running the GA with changing initial population sizes at a constant number of generations. The initial population size that gave us the least error was then taken which in this case is 20. After this the optimal initial population size was made constant whiles the number of generations were

varied. The number of generations that gave the least error was the taken as the optimal generations.

4.6 The Model

The model describes the basic flow of the algorithm for developing an ensemble of 9 classifiers from the 60 classifiers, see figure 3.1. The method of voting was then applied on the 9 chosen classifiers for generating the classification accuracy of the ensemble.

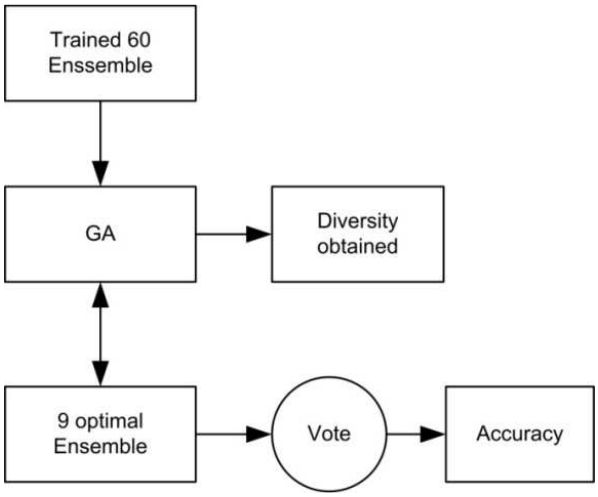

Figure 3.1: The mapping process of diversity and accuracy

4.7 Implementation

A vector of classifiers was created which was composed of 60 classifiers. This was because the more the classifiers there was, the better the search space for the GA for an optimal solution. All the classifiers in the vector were trained. The GA only looked for a solution for an ensemble of 9 classifiers. An odd number for the ensemble was chosen solemnly to avoid a tie when the method of voting was used. See figure 3.1 for the flow diagram for the system. The evaluation function was composed of two variables, the diversity measure and the targeted diversity (T). This diversity value was incremented from zero to 0.2. The zero would imply no diversity and

the 0.2 the increased diversity. The 0.2 was used because it was observed from the simulations that the highest diversity was around 0.2024. See equation (3.5) for the evaluation function used.

$$f_{GA} = -(kw_r - T_{kwr})^2 \tag{3.5}$$

Where: f_{GA} is the evaluation function, kw_r is the particular, diversity of the 9 classifiers and T_{kwr} is the targeted diversity.

The function is parabolic in the negative axis so that the optimal point is achieved when diversity measured is the same as the targeted diversity. The GA was then optimized by first searching the KWR values which the GA could easily approximate.

4.7.1 Vector of Classifiers

The classifiers were created via the normal distribution by creating them arbitrarily, hence the activation functions, hidden nodes, and the learning rate were chosen arbitrarily. This was so that the vector contained an ensemble of classifiers would not be biased. However a precaution was taken so that weak classifiers were not created, all the classifiers had the number of hidden nodes larger than the number of inputs. The vector also had classifiers that had a classification mean square error of less than 0.45 on the validation dataset. The ensemble of 60 vectors was optimized by using an ensemble that produced a greater diversity measure. This diversity measure is 0.2024. Intuitively this would be able to provide the GA with better classifiers that could generate the required diversity measure (KWR).

4.7.2 The Nine Ensemble of Classifiers

The validation dataset was used to select the nine classifiers from the vector of 60 classifiers. The classifiers were decrypted into a set of binary numbers as stated before. This binary number represented the ID of the individual classifiers; the bits can be seen from the individual columns. See table 3.2 for only the 9 classifiers.

Table 3.2: The IDS of the 9 classifiers

C1	C2	C3	C4	C5	C6	C7	C8	C9
1	1	1	1	1	1	1	1	1
0	0	0	1	0	0	0	0	0
1	0	1	0	1	1	0	1	1
1	0	0	0	1	0	1	0	0
1	0	0	1	1	1	0	0	0
0	1	1	0	1	1	1	0	0
1	0	0	1	0	0	1	0	0
0	1	0	0	1	0	0	1	0
0	0	1	0	0	1	0	0	1
0	0	0	0	0	0	0	0	0
0	1	0	1	0	0	1	1	0
1	1	1	1	1	1	0	0	1

The maximum diversity given by the ensemble of 60 classifiers was 0.2024; hence also the GA could not find any KWR value beyond this point. This further limited the number points that could be used to map the relationship between structural diversity and accuracy.

4.8 Results

Figure 3.2 and 3.3 show the results that were found from using the validation dataset. The ensemble of 9 classifiers chosen by the GA was then tested on the testing dataset so as to bring more sense to the results, see table 3.3.

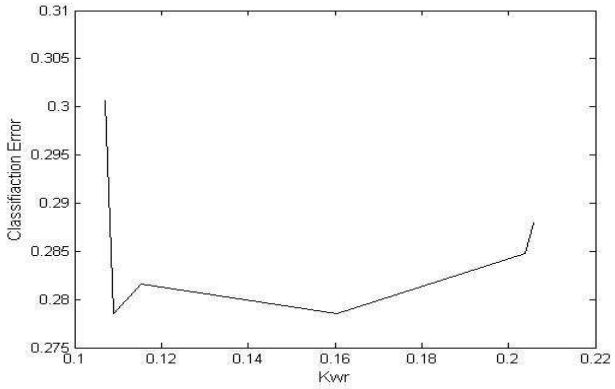

Figure 3.2: GA predicting 6 diversity values

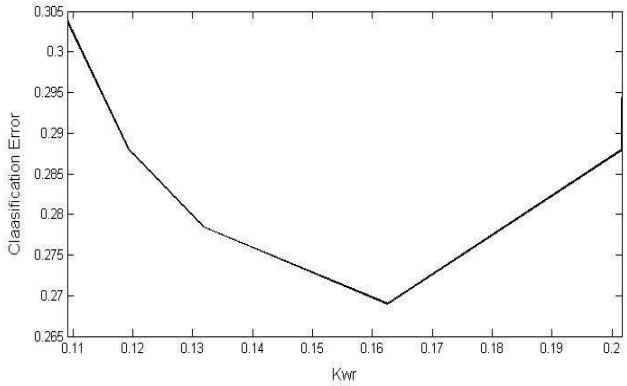

Figure 3.3: Optimized GA on the same 60 ensemble

Table 3.3: Classification error on the testing dataset

Kwr	Error (Initial *Kwr*)	Errors (Optimized *Kwr*)
0.11	0.3128	0.2821
0.16	0.2821	0.2749

It can be seen that the results from Figure 3 and Figure 4 follow the expected trend. The accuracy increases with increasing diversity. However there is a point where the degree of diversity becomes unfavorable. The accuracy began to drop with an increase in diversity. This is in alignment with [3], who stated that diversity can either profit the system or it could bring about poor performance on the classification. It can also be observed from the graphs that the data

points of interest are not to scale. The occurrence of a change is not consistent. This could be attributed to:

- The fact that there was a lot of rounding off values in the software package (Matlab),

- The other factor is that the ensemble of 60 classifiers was not designed with a linear or with consistent increments of diversity values.

- The targeted diversity values might not have been possible to be extracted from the ensemble and due to that the GA will provide its local solution.

Mean square error was used in all instances as a reference so as to observe the behavior of the ensemble classification with the measured changing structural diversity. This error is just a relative measure between the different ensemble classifiers that were used and these relative measures on the testing dataset showed that structural diversity can be used to measure the potential of improvement on the ensemble of classifiers since a relationship was observed between different structurally diverse ensembles and their classification.

4.9 Conclusion and Discussion

The aim of this study was to measure structural diversity of the ensemble, that means a measure on the diversity of the ensemble as seen from the structure of the individual classifiers and not seen from the outcomes of the individual classifiers and then to gain some insight on diversity and accuracy. This is necessary so that knowledge on whether diversity can be used to measure the potential for improvement of an ensemble of classifiers can be gained. The results show that there is a relationship between structural diversity and accuracy. As diversity increases the generalization ability of the ensemble improved. However it was observed that too much diversity increased the classification error. This study has also shown that diversity of an ensemble can be induced by having an ensemble that is composed of classifiers that have different parameters such as activation functions, number of hidden nodes and the learning rate, as was proposed by [28].

The methods used were computationally expensive since they made use of the GA and the training of 60 classifiers. This study agrees with most literatures that diversity does improve the accuracy of the ensemble [3]. This was observed by using the testing dataset on the ensemble that had a low classification error. This study was limited by the bank of classifiers (60 classifiers) that were created arbitrarily. This ensemble had 0.2024 diversity measures which meant that only small samples could be used to verify the relationship between diversity and accuracy.

Chapter 5

Investigating Ensemble weight Distributions for Indicating Structural Diversity

5.1 Introduction

This study analysed the distributions of the weight vectors (weights and biases) of the Multi-Layer Perceptron (MLP) composed within the ensemble. In particular the distribution parameter of concern is the variance which also leads to knowledge of the standard deviation of the weight vector samples from the mean. This was inspired by the fact that the number of hidden nodes controls the complexity of the Neural Network (NN), which translates to the structure of the (MLP). Aimed objectives of viewing ensemble diversity in terms of the structure and not the outcomes is to: broaden the research scope for ensemble diversity measures, add new understanding to ensemble diversity and possibly lead to other measures not focused on the classifiers' outcomes to measure ensemble diversity and lead to a unique definition of ensemble diversity hence robust measures for ensemble diversity.

The rest of the chapter attempts to meet these objectives. A method of voting was used to fuse or aggregate the individual classifiers for a final decision of the ensemble. The interstate conflict data was used for demonstrating the concept of structural diversity in this chapter. The sections on this chapter are organized as follows; the following section deals with ensemble diversity induction, structural diversity, methodology, results and discussion as well as the conclusion.

5.2 Ensemble Diversity

Two methods for inducing ensemble diversity of the classifiers are conducted. The first method uses the bagging algorithm, short for boostrap aggregation, and the second method is by using

classifiers with different number of hidden nodes, this method is one of the bases of the concept of Ensemble Structural Diversity (ESD). Bagging is considered to be one of the earliest ensemble algorithms and one of the simplest to implement [59]. The bagging algorithm trains the ensemble of classifiers by exposing the individual classifiers to a randomly chosen sample from the training data. These classifiers are normally known as weak or base learners, since they are trained to have classification accuracies of just above 50% [1]. The classifiers then can learn different domains of the problem and hence diversity can be induced.

Structural diversity would also imply that the classifiers have variations in the distributions of the classifier weights. The weight vector samples of the MLP are initialized from a Gaussian distribution and hence it will be expected that initially the weight vector samples have the same variance and the mean. This would mean that it is the training scheme of the ensemble that greatly influences the distribution parameters of the weight vector samples of the MLP. Hence using different complexities will also affect the weight vector samples differently and hence a study on the distribution of the weight vector might shed light on how the parameters of the distributions relate to ensemble generalization.

However, within the ensemble diversity studies, it is not only the measures of diversity that are of concern but also the aggregation methods. Shipp and Kuncheva [17], among many, have looked at the relationship between combination methods and measures of diversity and have found that certain measures of diversity correlated with certain aggregation schemes. It was also noted that the correlation observed had strong dependency on the data used [17]. This shows the complexity, of the ensemble diversity studies, which have also been noted by [60, 42, 3, 61], et al. This complexity is also expected for the ESD measures. This implies that developing a good measure and having a good aggregation scheme does not normally go hand in hand. However, for the sake of proof of concept only the majority vote scheme is considered due to its wide use.

5.2 Methodology

From error diversity measures, it was found that diversity reduces the variance in the decomposition of the error measure [60] and improves ensemble generalization. Two methods

are used to induce diversity, the first one is using a training algorithm (bagging) and the second one is having a committee of classifiers with different number of hidden nodes. Then a measure of diversity that looks at the variances of the classifier weights vector (weights and biases) is conducted. This measure is then compared to the generalization performance of the ensembles. Figure 5.1 shows the steps taken for the process mentioned. Only an ensemble containing five classifiers was used. Five classifiers were only considered so that computational cost was minimized and the number was made odd so that there were no ties during voting process, for the final decision. The classifiers initially had 8 hidden nodes with a linear activation function. However, when inducing diversity via the variation architecture of the classifiers, the number of hidden nodes was varied randomly between 8 and 21. This was so that they were not biased.

Certainty or confidence measures on the individual outcomes of the classifiers were done so that more knowledge could be gained on the generalization of the ensembles. The variance of certainties from the five classifiers will be used to give a more precise indication of the diversity of the ensemble. This means if there is no variance then the ensemble is highly certain which means that it could be highly biased. The certainties are measured from the outputs of the individual classifiers for the outputs of the classifiers are taken as a probability measure. Normally in a binary classification problem a 0.5 output is treated as a 1.

However, in this chapter a 0.5 output was not immediately rounded off to a 1, since its confidence measure is a zero, as can be seen from equation (5.1). The 0.5 output is taken as being the same as tossing a coin and thus its final outcome would either be a 0 or a 1. The certainty is symmetrical about the 0.5 output from the classifier. To illustrate equation (5.1), a 0.9 output from a classifier will have a confidence of 0.8 and a final classification output of 1. An output of 0.4 will be assigned a confidence of 0.2 and would mean a 0 for the final outcome (after rounding off). This equation is inspired from the Dynamically Averaging Networks (DAN) by Jimenez [62] and it was modified in this chapter.

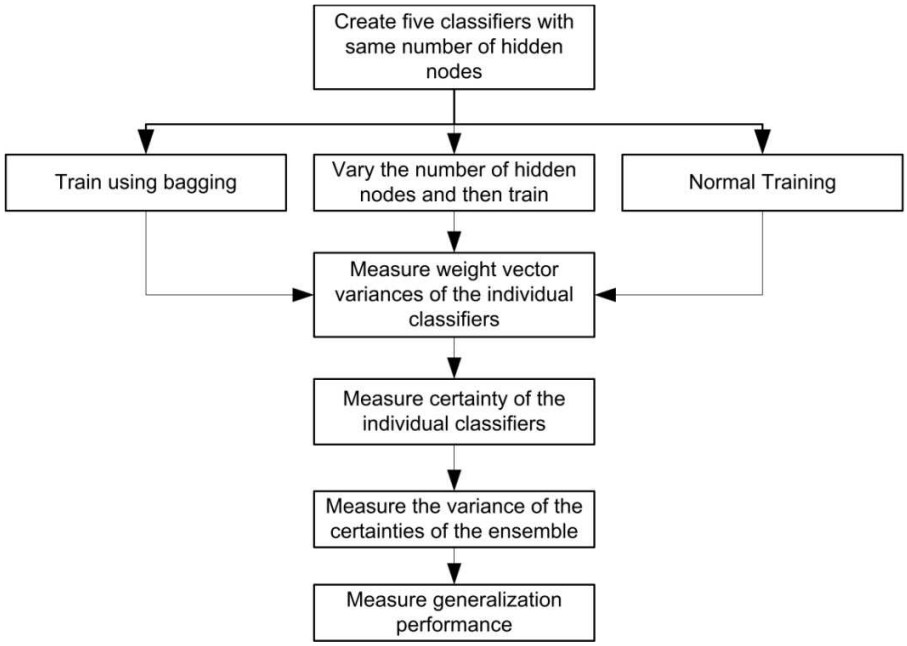

Figure 5.1: Flow diagram on diversity analysis

$$C\big(f(x_i)\big) = |2f(x_i) - 1| \tag{5.1}$$

Where, $f(x_i)$ is the immediate output from a classifier with input x_i before being processed into binary and $C\big(f(x_i)\big)$ is the certainty of the processed input data to a certain class [31].

$$S(x) = \frac{\sum_{f(x_i)=y_i} C(f_i)}{\sum C(f)} \tag{5.2}$$

where, S is the sum of all the classifiers which won the vote for an input data x_i, y_i is the correct output value from the data and the $\sum C(f)$ is the normalizing factor. The last step is to calculate the variance of the S vector for all the data samples that were correctly classified. This is the variance of the distribution of the certainties of the correctly classified data samples. See equation (5.2) for the calculation of the certainty distribution of the classifiers that won the vote. This certainty, say $S(x_i)$, is the confidence of the overall classifiers for that particular data sample. The certainties were normalized between 0 and 1, representing high to low certainties, respectively. The tests were done on the interstate conflict data [63].

The dataset has 7 features and a binary output. A 0 represents a conflict and a 1 represents peace. The data was conditioned such that it had approximately 50/50 conflict and peace cases. The training data was composed of 1006 and 869 for training and testing, respectively. The data was normalized between 0 and 1 so as to have same weighting for all the input features. An ensemble would be considered diverse if it had different variances on the distributions of the weights vector (weights and biases) between classifiers. A Multi Layered Perceptron (MLP) was used for all the experimentation. The certainty measures were conducted on the test dataset where else the weights vector variance measures, were conducted after the classifiers were trained.

5.3 Results and Discussion

Two methods of inducing diversity have been studied, the bagging and parameter change of the ensemble of classifiers. The structural diversity was observed over the two diverse ensembles. The results on the accuracies do not show any difference between the different ensembles. They all produced accuracies of approximately 74 %. However, it is evident from the variances of the weights vector that the ensemble is diverse. The variances on the weight vector due to the bagging algorithm were still on a close range, see Table 5.1. This showed that less diversity was induced on this ensemble which could be due to the use of strong learners.

However, the variances due to structural diversity (changing of hidden nodes) produced classifiers that had the vector weight variances significantly different within the committee, see Table 5.1 on the second column. Intuitively one would conclude that this ensemble was more diverse as compared to the ensemble trained via bagging. But the bagged ensemble produced better generalization performance as compared to the other ensembles (diverse ensemble due to different structural parameters and the no-diverse ensemble), (see Table 5.2). This might mean that by observing the vector weights as a measure of ensemble diversity might not relate well with the generalization performance of the ensemble. This therefore leads to conclude that the measure of structural diversity in this regards is insignificant when related to the classification accuracy.

The significant weight vector variances on the ensemble (among the classifiers) with different number of hidden nodes can be attributed to the use of different number of hidden nodes. This then shows that the use of the weight vector distributions might not be a good method to correlate diversity and generalization. This is one of the biggest disadvantages of measuring diversity from the structural point of view. For one can develop a good measure but then lose out on using the measure to predict the generalization performance of the ensemble, as noted. According to these results, see Table 5.1 and 5.2, when the ensemble is non-diverse then the variance of the certainties is zero. This means the non-diverse ensemble is extremely certain and there is no variation. Intuitively, this made sense for an ensemble that is non-diverse, for it would be biased. This means that this certainty measure should not be confused with the confidence of reducing risk in classification.

Table 5.1: Variances of the diverse and non diverse ensembles

Bagged (σ^2)	**Nodes (σ^2)**	**Non-Diverse (σ^2)**
0.25915	0.4582	0.26391
0.30675	0.23119	0.26391
0.27167	0.22072	0.26391
0.23999	0.18754	0.26391
0.29347	0.16668	0.26391

Table 5.2: Accuracies and variance measures

Acc (Bagged)	**74.914**
Acc (Nodes)	74.569
Acc(non-diverse)	74.338
$\sigma 2(C_{div}(f(x))$	0.0064141
$\sigma^2(C_{non\text{-}div}(f(x))$	0

The confidence in this context measured the extent to which the individual classifiers believed to have been correct not necessary that the classification was correct. This shows how structural diversity measures can better bring understanding to the classification problems. These results show that the data used was not complex enough and one classifier would be adequate for this problem for the ensemble produced approximately similar generalization performance. This is concurrent with literature that diversity can both be harmful or beneficial [38]. Further work can be done by using different aggregation schemes even the Dynamic Average Networks (DAN) for

understanding the structural variation of the classifiers. Structural diversity measures could also be attempted in other artificial machines.

5.4 Conclusion

This chapter presented concepts inspired from statistical methods and certainty to better understand the structural diversity of the ensemble. A different assessment of ensemble diversity has been presented as opposed to looking at the classifier outcomes to measure ensemble diversity. The weight vector of the classifiers has been assessed for indicating ensemble diversity. Due to looking at the outcomes and not really at the structure, misleading judgments about the ensemble diversity might be taken and hence poor correlation judgments on the generalization of the ensemble and ensemble diversity.

This was observed from having a structurally diverse ensemble having almost the same generalization as the non-diverse ensemble. However, this could be attributed to the size of the data and the data used. Certainty or confidence variance of zero was recorded due to a non-diverse ensemble which indicated that the non-diverse ensemble was biased. The consideration of viewing ensemble diversity from structural point of view (from the learning parameters) adds more knowledge on the distributions of the vector weights of the diverse ensembles and hence on the ensemble diversity research community. Measuring ensemble diversity from the vector of weights might be meaningful but correlating this measure with ensemble generalization might not be meaningful. A formal definition of ensemble diversity still remains an open discussion. More work can still be done in changing the number of the classifiers within a committee and by using other datasets. Classifiers have a number of parameters depending on the artificial machine used, which leaves the search for other methods of measuring structural diversity for exploration.

Chapter 6

Conclusion

6.1 Summary of Findings

The aim of this research was to introduce and investigate a new way of viewing ensemble diversity, called the ensemble structural diversity. The investigation was conducted by inducing structural diversity within the ensemble and then measuring this structural diversity. This enabled the use of GAs resulting in a relationship between structural diversity and generalization performance being captured. Different ensemble structural diversity measures were conducted, the ecological measures and the statistical measure indicated that as the structural diversity increased so did the generalization performance (accuracy) of the ensemble. However, there was a point where ensemble structural diversity became unfavourable, as ensemble structural diversity increased the generalization performance decreased. However, it was observed that this point was different for different ensemble structural diversity measures used and hence further research on the behaviour of this point were beyond the scope of this dissertation. These results indicate that structural diversity serves as a potential for use in constructing efficient and effective ensembles. The first problem that was encountered was that of representation, how one represents the structures that compose the MLP into variables that could be measured, so that an ensemble of MLPs with different structural parameters can be quantified. This was successfully attained and a good relationship between ensemble structural diversity and generalization performance was observed for as the diversity increased so did the generalization performance of the ensemble, however, to a certain region of the diversity values. Through the measures conducted, the classifiers were for the first time seen as ecological species within the ensemble. This was done so that ecological methods of measuring biodiversity could be implemented. In this way, structural diversity was able to be quantified and hence could be compared to the generalization performance of the ensemble. A similar work was conducted with the use of a

different dataset and the results were concurrent [26], with the work done. Shannon was also interpreted from the information theory side and was also used as the uncertainty measure. This uncertainty measure was inspired from information theory and was taken as the diversity measure. The uncertainty in this context means less biasing on the ensemble indicating a diverse ensemble. The results indicated that as the uncertainty increased so did the generalization performance of the ensemble.

The last method for measuring ensemble structural diversity was by observing the distributions of the classifiers weight vector (weights and biases). This was because a comparison was conducted between the induction of diversity by structure and by a training algorithm (bagging). This also indicated the existence of ensemble structural diversity which was, however, not that different from the non-diverse ensemble. This was because this measure of ensemble diversity was mainly concerned with indicating ensemble diversity and not specialized to relate ensemble diversity with generalization performance. In this regard, it was considered the weakest form among the other studies conducted of relating ensemble diversity with generalization performance with the aim of validating the measure.

This research has shown that diversity within the ensemble is important even when measured from the structural parameters and as opposed to have been measured from the outcomes as has been commonly practiced. Ensemble diversity can be induced via the structural parameters that constitute the individual classifiers. Further, as opposed to looking at the outcomes as a means of measuring ensemble diversity, ensemble diversity can be measured from the structural view point. This is achieved by looking at the structural parameters composing the classifiers, hidden nodes, activation functions and the learning rate. Chapter 3 and 4 induce ensemble diversity by structure where else in chapter 5 diversity induced by structure and by a learning algorithm.

It was found that the results were concurrent with literature on the basis of the necessity of diversity for efficient ensembles. Even though this might bring about a good measure of ensemble diversity, it was found in some instances that the measure of structural diversity did not correlate well with the generalization performance of the ensemble. This was because one can have a structurally diverse ensemble where else all the classifiers composed within the ensemble

approximate the same input to output function. This is where the measure of looking at ensemble diversity from the outcome outweighs that of being viewed from structure. Hence, the use of structural diversity with the aim of building efficient ensembles is not strongly encouraged even though it has shown to provide a possibility for constructing efficient ensembles. This study showed that diversity can be measured differently and this is due to the fact that it has no formal definition. Only one dataset was used as a form of proving the concept (structural diversity) by using different measures. Hence, due to the lack of the use of other datasets this research still remains to be explored and hence no conclusion could be done on whether the measures conducted are robust for all situations and representations of other datasets. However, we can conclude that structural diversity does exist and the research community can begin to revisit their ensemble diversity measures as a form of assessing the predictive ability of the ensembles.

6.2 Recommendations and Future Work

It is recommended that since diversity has no formal definition, a hybrid ensemble to ensemble construction may be used. This means that as opposed to the use of parameters to measure diversity, one would look at the different machines constituting the ensemble as a means of measuring diversity. Future work might entail the use of other datasets onto this concept of ensemble structural diversity. There seems to be many variables within ensemble systems and all these variables play a major role in the relationship between ensemble diversity and generalization performance. Ensemble systems entail, the training, number of classifiers used to make up the ensemble, the complexity of the classifiers (hidden notes, learning rate and activation function) , the aggregation scheme for a final solution and even the data used to some certain extent. It is until these parameters are well related to one another that a robust generic measure of ensemble diversity can be developed. More study on the point where diversity becomes unfavorable for the generalization of the ensemble should be conducted.

Appendix A

Ensemble Systems

An ensemble system is whereby one uses more than one artificial machine to do a certain task. However the committee of the classifiers in this context need to be diverse so that the system is not composed of a number of classifiers whereby if one classifier was used, it could have performed the same way as the use of a committee, due to lack of diversity within the committee. Diversity can be induced either be by structure, different training scheme, different feature subsets and combination of different classifiers. The goal of these different methodologies is so that the classifiers within the ensemble have different decision boundaries, see figure A.1, which was taken from Polikar [1]. These classifiers are then combined so as to produce a strong classifier as compared to the use of one classifier, see figure A.1, The sum function can either be a majority, averaging scheme, etc.

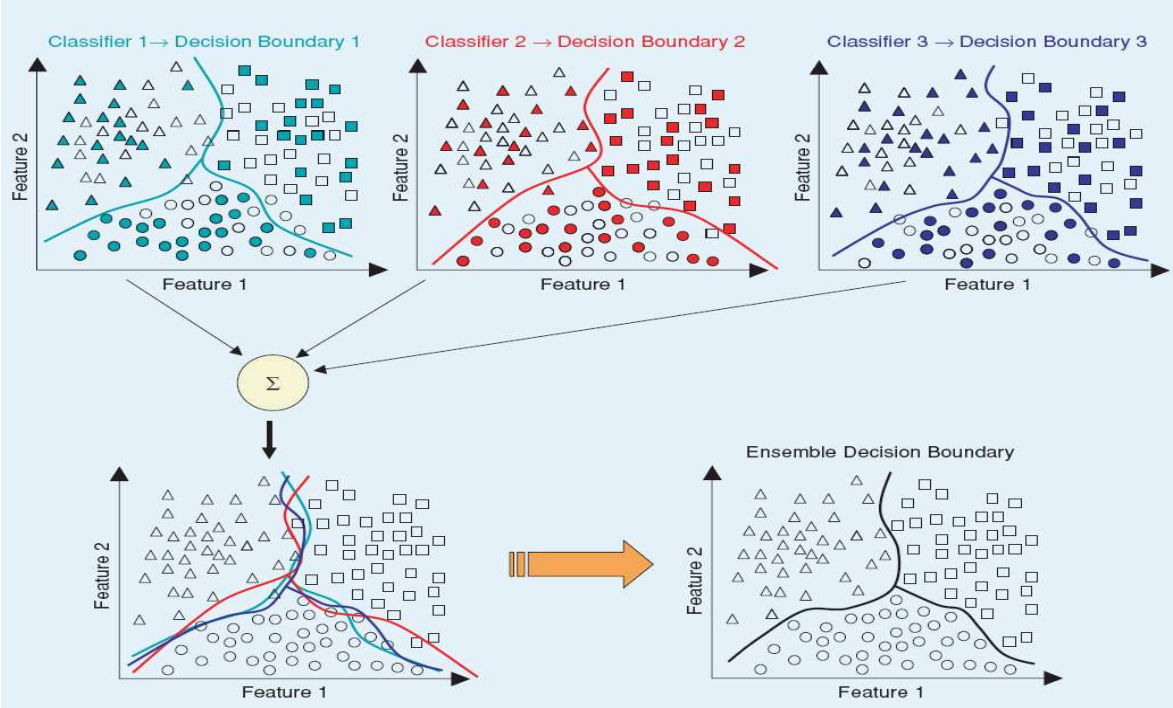

Figure A1: Combining classifiers with different decision boundaries [1].

A.1 Ensemble diversity measures

The research community has sort to develop ensemble diversity measures. However many of the researchers have considered diversity measures at the outcomes and not by structure. At the outcomes is when one looks at the decision boundary of the individual classifiers and then compare the variations or the differences that exists within the ensemble so as to produce a quantity. This quantity would be the measured diversity of the ensemble. A list of the ensemble diversity measures that measure diversity from the outcomes are listed in the following sections.

A.1.1 Measures of Diversity

In order to quantify outcome diversity, several measures have been defined. These are pair-wise measures and non pair-wise measures. All these measures were taken from Polikar [1], Kuncheva and Whitaker [4]. Pair-wise measures are defined between two classifiers.

-***Pair wise measures*** [1, 4],

Table A.1: A matrix showing the relationship between a pair of classifiers.

	$D_k\ correct(1)$	$D_k\ wrong(0)$
$D_i\ correct(1)$	N^{11}	N^{10}
$D_i wrong(0)$	N^{01}	N^{00}

$$Total\ N = N^{11} + N^{01} + N^{00} + N^{10}$$

Q Statistics

This diversity measure is used to evaluate the degree of similarity and dissimilarity in the outcomes of the classifiers within the ensemble. This measure of diversity focuses on the outcome of a pair of classifiers. It is however essential to measure the degree of the agreement

and disagreement on the outcomes of the ensemble to ensure the limitation of the outcomes. For classifiers i and j the outcome based diversity is given by (A1.1)

$$Q_{i,j} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \quad \text{(A1.1)}$$

where:

N^{11}- represents cases where both classifiers correctly classified instances

N^{00}- represents misclassification of instances

N^{01}- represents cases in which i misclassified an instance whilst classier j correctly classified that particular instance.

The averaged values of Q over all pairs in the ensemble is,

$$Q_{ave} = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{k=i+1}^{L} Q_{i,j} \quad \text{(A1.2)}$$

- The range of the index varies between -1 and 1.
- The maximum diversity is obtained at 0.
- Classifiers that recognize similar objects have positive Q values
- Classifiers that commit errors on different objects result in a negative Q

Correlation measure

This measure of diversity looks at the correlation between two classifiers outputs. This uses table 1 as in Q statistics. The diversity is defined by (A.1.3)

$$p_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{\sqrt{(N^{11}+N^{10})(N^{00}+N^{01})(N^{11}+N^{01})(N^{10}+N^{00})}} \quad \text{(A.1.3)}$$

- The range of the index varies between -1 and 1
- Maximum diversity is obtained at p = 0.

- Negative p mean classifiers commit errors at different places
- Positive p mean classifiers recognize similar objects correctly.

The disagreement measure

This measure is the probability that two classifiers will disagree. It is the ratio between the number of observation on which one classifier is correct and the other is incorrect to a total number of observations.

$$Dis_{i,k} = \frac{N^{01}+N^{10}}{N^{11}+N^{10}+N^{00}+N^{01}} \quad \text{(A.1.4)}$$

- Diversity increases with increasing the disagreement

The double fault measure

It is defined as the proportion of the cases that have been misclassified by both classifiers, see equation (A.1.5)

$$DF_{i,k} = \frac{N^{00}}{N^{11}+N^{10}+N^{00}+N^{01}} \quad \text{(A.1.5)}$$

- Diversity increases with increasing the double fault

-***Non Pair-wise measures***[1, 4]

Entropy Measure

This measure makes an assumption that the diversity is highest when half of the classifiers are correct and the other half is incorrect. The entropy measure is defined as:

$$E = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T-[T/2]}min\{\delta_i, (T-\delta_i)\}$$

Where:

δ_i – The classifiers that misclassifies instance x_i

T – The total number of classifiers

N – The total number of data samples

- The entropy diversity measure varies between 0 and 1,
- 0 means the classifiers are practically the same and 1 means they are different.
- This means that 0 means lowest diversity and 1 means highest diversity.

Kohavi-Wolpert Variance

This measure follows the similar approach to the disagreement measure.

$$KW = \frac{1}{NT^2}\sum_{i=1}^{N} \delta_i \,.\, (T - \delta_i) \qquad \text{(A.1.6)}$$

where:

δ_i – The classifiers that misclassifies instance x_i

T – The total number of classifiers

N – The total number of data samples

KW – The diversity index.

Measure of difficulty

This measure uses the random variable that is defined as the fraction of classifiers that misclassifies x_i. The measure of variance is then the variance of the random variable Z, see equation (A.1.7).

$$\theta = \frac{1}{T}\sum_{i=1}^{T}(z_t - \bar{z})^2 \qquad \text{(A.1.7)}$$

where $\bar{z}$ - is the mean of z., hence is the average fraction of classifiers that misclassifies any given input.

T – The total number of classifiers

θ-diversity index

The generalized diversity

This measure has been proposed by Partridge [64]. It is defined by:

$$p(1) = \sum_{i=1}^{L} \frac{i}{L} p_i$$

$$p(2) = \sum_{i=1}^{L} \frac{i(i-1)}{L(L-1)} p_i$$

where p_i- The probability that a random variance expressing the proportion of classifiers that are incorrect.

$L-$ The total number of classifiers

The generalized diversity equation is given by

$$GD = 1 - \frac{p(2)}{p(1)}$$

- GD varies between 0 and 1.
- GD = 0 (minimum diversity when $p(2) = p(1)$)
- GD = 1 (maximum diversity when $p(2) = 0$)

The confidence measure

This is similar to GD.

$$CFD = \begin{cases} 0, & p_0 = 1.0; \\ \dfrac{1}{1-p_0}\displaystyle\sum_{i=1}^{L}\dfrac{L-i}{L-1}p_i, & p_{0<1} \end{cases}$$

- Minimum diversity value is 0, when all classifiers are always correct or when all the classifiers are simultaneously either correct or wrong.
- Maximum value is achieved at diversity of 1 when all the misclassifications are unique, when at most one classifier will fail on any randomly chosen object.

A.2 Bagging

This is one of the well known training algorithms that were used in this study [1]. This algorithm ensures continuous resample of the training data inducing a diverse ensemble. For in the process weak and strong classifiers are developed, taking place within the same dataset. Once the training is done the final decision of the ensemble is taken by the use of a majority vote. Here is the algorithm for the bagging training process which was also taken from Polikar [1].

Input:

- Training data S with correct labels $w_i \in \Omega = \{w_1, \dots, w_c\}$ representing C classes
- Weak learning algorithm Weak-Learn,
- Integer T specifying number of iterations
- Percent (or fraction) F to create bootstrapped training data

Do $t = 1, \dots, T$

1. Take a bootstrapped replica S_t by randomly drawing F percent of S.
2. Call Weak-Learn with S_t and receive the hypothesis (classifier) h_t
3. Add h_t to the ensemble, E

End

Test: Simple Majority Voting- Given unlabeled instances **x**

1. Evaluate ensemble $E = \{h_t, \dots, h_T\}$ on **x**

2. Let $v_{t,j} = \begin{cases} 1, & if\ h_t\ picks\ class\ w_t \\ 0, & otherwise \end{cases}$

   be the vote given to class w_j by classifier h_t.

3. Obtain total vote received by each class

   $$V_j = \sum_{t=1}^{T} v_{t,j}, j = 1, \dots, C$$

4. Choose the class that receives the highest total vote as the final classification

References

[1] R. Polikar, "Essemble based systems in decision making," *Circuits and Systems Magazine, IEEE,* vol. 6, pp. 21-45, 2006.

[2] L. I. Kuncheva and S. T. Hadjitodorov, "Using diversity in cluster ensembles," *Systems, Man and Cybernetics, 2004 IEEE International Conference on,* vol. 2, pp. 1214-1219 vol.2, 2004.

[3] L. I. Kuncheva, M. Skurichina and R. P. W. Duin, "An experimental study on diversity for bagging and boosting with linear classifiers," *Information Fusion,* vol. 3, pp. 245-258, 12. 2002.

[4] L. I. Kuncheva and C. J. Whitaker, "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy," *Machine Learning,* vol. 51, pp. 181-207, 2003.

[5] B. V. Dasarathy and B. V. Sheela, "A composite classifier system design: Concepts and methodology," *Proc IEEE,* vol. 67, pp. 708-713, 1979.

[6] L.K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 12, pp. 993-1001, 1990.

[7] G. Brown, J. Wyatt, R. Harris and X. Yao, "Diversity creation methods: a survey and categorisation," *Information Fusion,* vol. 6, pp. 5-20, 3. 2005.

[8] B. Igelnik, Y. H. Pao, S. R. Leclair and C. Y. Shen, "The ensemble approach to neural-network learning and generalization," *IEEE Trans. Neural Netw.,* vol. 10, pp. 19-30, 1999.

[9] J. Kittler, M. Hatef, R. Matas and J. Duin, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 20, pp. 226-239, 1998.

[10] L. I. Kuncheva, "Combining Pattern Classifiers: Methods and Algorithms (Kuncheva, L.I.; 2004) [book review]," *Neural Networks, IEEE Transactions on,* vol. 18, pp. 964-964, 2007.

[11] T. G. Dietterich, "Machine-Learning Research: Four Current Directions," *The AI Magazine,* vol. 18, pp. 97-136, 1998.

[12] R. Polikar, D. Parikh and S. Mandayam, "Multiple classifier systems for multisensor data fusion," *Sensors Applications Symposium, 2006. Proceedings of the 2006 IEEE,* pp. 180-184, 2006.

[13] Sharkey, A. J. C. and N. E. Sharkey, "Combining diverse neural nets," *The Knowledge Engineering Review,* vol. 12, pp. 231-247, 1997.

[14] L. I. Kuncheva, "A theoretical study on six classifier fusion strategies," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 24, pp. 281-286, 2002.

[15] D. Parikh and R. Polikar, "An Ensemble-Based Incremental Learning Approach to Data Fusion," *Systems, Man, and Cybernetics, Part B, IEEE Transactions on,* vol. 37, pp. 437-450, 2007.

[16] D. Ruta and B. Gabrys, "Classifier selection for majority voting," *Information Fusion,* vol. 6, pp. 63-81, 2005.

[17] C. A. Shipp and L. I. Kuncheva, "Relationships between combination methods and measures of diversity in combining classifiers," *Information Fusion,* vol. 3, pp. 135-148, 6. 2002.

[18] H. Bunke and A. Kandel, Hybrid Methods in Pattern Recognition, World Scientific, pp. 199-226, 2002.

[19] K. Sirlantzis, S. Hoque and M. C. Fairhurst, "Diversity in multiple classifier ensembles based on binary feature quantisation with application to face recognition," *Applied Soft Computing,* vol. 8, pp. 437-445, 1. 2008.

[20] L. I. Kuncheva and C. J. Whitaker, "Ten measures of diversity in classifier ensembles: limits for two classifiers," *Intelligent Sensor Processing (Ref. no. 2001/050), A DERA/IEE Workshop on,* pp. 1001-1010, 2001.

[21] S. Ridella and R. Zunino, "Empirical measure of multiclass generalization performance: the K-winner machine case," *IEEE Trans. Neural Netw.,* vol. 12, pp. 1525-1529, 2001.

[22] H. Zouari, L. Heutte and Y. Lecourtier, "Controlling the diversity in classifier ensembles through a measure of agreement," *Pattern Recognition,* vol. 38, pp. 2195-2199, 11. 2005.

[23] K. M. Ali and M. J. Pazzani, "Error reduction through learning multiple descriptions," *Machine Learning,* vol. 24, pp. 173-202, 1996.

[24] L. K. Hansen and P. Salamon, "Neural network ensembles," in *IEEE Transaction on Partten Analysis and Machine Intelligence,* vol. 12, pp. 993-1001, 1999.

[25] L. I. Kuncheva and R. K. Kountchev, "Generating classifier outputs of fixed accuracy and diversity," *Pattern Recognition Letters,* vol. 23, pp. 593-600, 3. 2002.

[26] R. Musehane, F. A. Netshiongolwe, L. Masisi, F. V. Nelwamondo and T. Marwala, "Relationship between Structural and Performance of Multiple Classifiers for Decision Support," *Proceedings of the, Nineteenth Anual Symposium of the Pattern Recognition Association of South Africa,* pp. 109-114, 2008.

[27] R. Polikar, L. Udpa, S. Udpa and V. Honavar, "An incremental learning algorithm with confidence estimation for automated identification of NDE signals," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control,* vol. 51, pp. 990-1001, Aug. 2004.

[28] A. Sharkey, "Combining artificial neural nets: ensemble and modular multi-net systems," *Connection Science,* pp. 1-30, 1999.

[29] B. Mirkin, *Clustering for Data Mining: A Data Recovery Approach.* ,First ed.London: Chapman and Hall/HRC, 2005.

[30] L. Masisi, F. V. Nelwamondo and T. Marwala, "The Effect of Structural Diversity of an Ensemble of Classifiers on Classification Accuracy," *International Association of Science and Technology for Development: Modeling and Simulation, IASTED,* pp. 135-140, 2008.

[31] L. M. Masisi, F. V. Nelwamondo and T. Marwala, "Investigating ensemble weight distributions for indicating structural diversity," in *15th International Conference on Neural Information Processing of the Asia-Pacific Neural Network Assembly* (accepted), 2009.

[32] L. Masisi, F. V. Nelwamondo and T. Marwala, "The use of entropy to measure ensemble structural diversity," *6th IEEE International Conference on Computational Cybernetics,* pp. 41-45, 2009.

[33] M. Kawakita, M. Minami, S. Eguchi and C. E. Lennert-Cody, "An introduction to the predictive technique AdaBoost with a comparison to generalized additive models," *Fisheries Research,* vol. 76, pp. 328-343, 12. 2005.

[34] X. Wang and H. Wang, "Classification by evolutionary ensembles," *Pattern Recognition,* vol. 39, pp. 595-607, 4. 2006.

[35] K. Tumer and J. Ghosh, "Linear and order statistics combiners for pattern classification," *Combining Artificial Neural Nets,* pp. 127-161, 1999.

[36] L. Nanni and A. Lumini, "Using ensemble of classifiers for predicting HIV protease cleavage sites in proteins," in *Amino Acids,* 2008.

[37] A. Assareh, M. H. Moradi and V. Esmaeili, "A novel ensemble strategy for classification of prostate cancer protein mass spectra," *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, vol. 2007, pp. 5988-5991, 2007.

[38] L. I. Kuncheva, C. J. Whitaker, C. A. Shipp and R. P. W. Duin, "Is independence good for combining classifiers?" *15th International Conference on Pattern Recognition*, vol. 2, pp. 168-171, 2000.

[39] G. Fumera, R. Fabio and S. Alessandra, "A theoretical analysis of bagging as a linear combination of classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, pp. 1293-1299, 2008.

[40] M. M. Islam, X. Yao, S. M. Shahriar Nirjon, M. A. Islam and K. Murase, "Bagging and boosting negatively correlated neural networks," *IEEE Trans. Syst. Man. Cybern. B. Cybern.*, vol. 38, pp. 771-784, 2008.

[41] T. Windeatt, "Diversity measures for multiple classifier system analysis and design," *Information Fusion*, vol. 6, pp. 21-36, 3. 2005.

[42] E. Cantu-Paz and C. Kamath, "An empirical comparison of combinations of evolutionary algorithms and neural networks for classification problems," *IEEE Trans. Syst. Man. Cybern. B. Cybern.*, vol. 35, pp. 915-927, 2005.

[43] C. M. Bishop, *Neural Networks for Pattern Recognition.* Oxford-London: Oxford University Press, 1995,

[44] K. M. Calder, M. J. Agnew, D. W. Stashuk and L. McLean, "Reliability of quantitative EMG analysis of the extensor carpi radialis muscle," *J. Neurosci. Methods*, vol. 168, pp. 483-493, 2008.

[45] J.L. Johrendt, P.R. Frise and M. A. Malik, "Streamlining automotive product development using neural networks," *International Journal of Vehicle Design*, vol. 47, pp. 19-36, 2008.

[46] N. Karunanithi, D. Whitley and Y. K. Malaiya, "Using neural networks in reliability prediction," *Software, IEEE*, vol. 9, pp. 53-59, 1992.

[47] K. Hung, Y. Cheung and L. Xu, "An extended ASLD trading system to enhance portfolio management," *IEEE Transactions on Neural Networks*, vol. 14, pp. 413-425, 2003.

[48] M. Pan, P. Li and Y. Cheng, "Remote online machine condition monitoring system," *Measurement*, vol. 41, pp. 912-921, 10. 2008.

[49] K. Tumer and J. Ghosh, "Analysis of decision boundaries in linearly combined neural classifiers," *Pattern Recognition,* vol. 29, pp. 341-348, 2. 1996.

[50] T. Heskes, "Bias/Variance Decompositions for Likelihood-Based Estimators," *Neural Comput.,* vol. 10, pp. 1425-1433, 1998.

[51] R. Kohavi and D.H. Wolpert, "Bias plus variance decomposition for zero-one loss functions," *Proceedings of Thirteenth International International Conference on Machine Learning,* pp. 275-283, 1996.

[52] J. H. Holland, "Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence," in Anonymous University of Michigan Press, 1975.

[53] A. H. Wright, "Genetic algorithms for real parameter optimization," *Foundations of Genetic Algorithms,* vol. 4, pp. 205-218, 1991.

[54] T. Tettey, "A Computational Intelligence Approach to Modelling Interstate Conflict: Forecasting and Causal Interpretations," Unpublished master's thesis, University of the Witwatersrand, South Africa/Johannesburg, 2007.

[55] T. D. Schneider, "Information Theory Primer," 2007.

[56] I. Harrison, M. Laverty and E. Sterling, "Species Diversity," Connexions Module: m12174, Version 1.3, Connexions Project, 2004.

[57] R. Kindt, A. Degrande, L. Turyomurugyendo, C. Mbosso, P. Van Damme and A.J. Simons, "Comparing species richness and evenness contributions to on farm tree diversity for data sets with varying sample sizes from Kenya, Uganda, Cameroon and Nigeria with randomized diversity profiles," *IUFRO Conference on Forest Biometry, Modelling and Information Science,* 2001.

[58] S. Baumgartner, "Why the measurement of species diversity requires prior value judgements," Department of Economics, University of Heidelberg, Heidelberg, 2006.

[59] L. Breiman, "Bagging predictors," Machine Learning, Vol. 24, pp. 123-140, 1996.

[60] G. Brown, "Diversity in neural network ensembles," PhD thesis, The University of Birmingham, 2004.

[61] S. Izrailev and D. K. Agrafiotis, "A method for quantifying and visualizing the diversity of QSAR models," *Journal of Molecular Graphics and Modelling*, vol. 22, pp. 275-284, 3. 2004.

[62] D. Jimenez, "Dynamically Weighted Ensemble Neural Networks for Classification," *Proceedings of the 1998 International Joint Conference on Neural Networks*, pp. 753-756, 1998.

[63] T. Marwala, "Bayesian training of neural networks using genetic programming," *Pattern Recogn. Lett.*, vol. 28, pp. 1452-1458, 2007.

[64] D. Partridge, "Network Generalization Differences Quantified," Neural Networks, Vol. 9, pp. 263-271, 1996.