ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 4, July 2014

Ensemble Techniques ToIncrease Classification Accuracy (ETICA)

Tiruveedula GopiKrishna

Abstract*—In this paper, Aggregation of predictions of multiple classifiers with the goal of improving accuracy has presented. An ensemble for classification is a composite model, made up of a combination of classifiers. The overall goal was to suggest a novel technique for ensemble creation, constantly producing accurate ensembles on most data sets. The basic approach used is to first create a number of models using Artificial Neural Networks and then use evolutionary algorithms to select and combine these models into an ensemble.Since there is no unified procedure to implement these steps, this paper proposes a new methodology to design Artificial Neural Network ensembles using a Genetic Algorithm to increase classification accuracy.

***Index Terms*—Artificial Neural Networks, Genetic Algorithms, Ensembles, Genetic Ensemble Member Selection.**

I. INTRODUCTION

In the first study, which was originally reported in [1], a simple algorithm based on GAs was evaluated against several straightforward ways of combining ANNs into ensembles. The suggested approach used GAs to search among all possible combinations of the available ANNs. The resulting ensemble is therefore just a subset of the available ANNs, here combined using averaging. The fitness function was based on ensemble accuracy on training and/or validation sets.

The second study, originally published in [2], introduced a novel technique for ensemble creation. The technique, named Genetic Ensemble Member Selection (GEMS), first trains a large number of ANNs (between 10 and 50) and then uses genetic programming to build the ensemble by combining available ANNs. The use of genetic programming makes it possible for GEMS to not only consider ensembles of very different sizes, but also to use ensembles as intermediate building blocks, which could be further combined into larger ensembles. The fitness function was again ensemble accuracy on training and validation sets. In this study, GEMS was only evaluated on four data sets, and the results were slightly discouraging. The main problem was that GEMS obtained extremely high accuracy on the parts of the data set used during evolution, but sometimes failed to generalize to test data.In these studies, altogether 28 publicly available data sets were used. The following 11 data sets were added to the ones used when evaluating Genetic-Regular Expression.

- **Ecoli:** A biological data set where the purpose is to predict the localization site of a protein.
- **Hepatitis Domain (Hepatitis):** Prediction of whether a patient will survive or not based on several medical measurements.
- **Horse colic database (Horse):** Prediction of whether a horse was operated on or not based on medical measurements.
- **Image:** The instances were drawn randomly from a database of 7 outdoor images. The images were hand segmented to create a classification for every pixel. Each instance is a 3x3 region, and the classes are brickface, sky, foliage, cement, window, path and grass.
- **LED Display (Led7):** This problem contains 7 Boolean attributes (a led segment on or off) and 10 classes, representing the set of decimal digits. The problem would be easy if not for the introduction of noise.
- **Satellite Image (Satellite):** The database consists of the multi-spectral values of pixels in 3x3 neighborhoods in a satellite image, and the classification associated with the central pixel in each neighborhood. The aim is to predict this classification, given the multi-spectral values.
- **Sick:** Prediction whether a patient is hyperthyroid or not based on medical measurements.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 4, July 2014

- **Thyroid:** This data set is very similar to Sick, and the target variable is again whether a patient is hyperthyroid or not. The Sick data set has more attributes but fewer instances, though.

- **Waveform:** This is an artificial three-class problem, based on three waveforms. Each class consists of a random convex combination of two waveforms sampled at the integers with noise added. A description for generating the data is given in [3].
- **StatLog vehicle silhouette (Vehicle):** This data set originated from the Turing Institute, Glasgow, Scotland. The problem is to classify a given silhouette as one of four types of vehicle, using a set of features extracted from the silhouette. The four vehicles are double Decker bus, Chevrolet van, Saab 9000 and Opel Manta 400.

- **Votes:** The task consists of classifying US Congressmen as democrat or republican based on their voting record on 16 key votes.

For a summary of the characteristics of these data sets, see **Table 1**. below.

Data sets	Instances	Classes	Continuous inputs	Categorical inputs
Ecoli	338	8	7	0
Hepatitis	156	2	6	13
Horse	367	2	7	14
Image	2314	7	19	0
Led7	3210	10	0	7
Satellite	6433	6	36	0
Sick	2887	2	7	22
Thyroid	3139	2	7	18
Waveform	5120	3	21	0
Vehicle	757	4	18	0
Votes	546	2	0	16

Table 1: UCI data set characteristics

II. BUILDING ENSEMBLES USING GAS

The overall purpose of this study was to introduce a simple, GA-based, technique for creating ensembles, and compare this to several straightforward alternatives. More specifically, altogether 18 alternatives, categorized in six groups, were evaluated.

A. Method

When conducting the experiments, each data set was divided in three parts; training, validation and test. The training set was used to train individual ANNs. The validation set was, as usual, not utilized during training, but it was intended to give an indication of the generalization capability. In this study, validation sets were used in different ways to rank and select ensembles. Setups not using any kind of selection used all data except the test set for training. Naturally, test sets were only used for the actual evaluation of each ensemble.10-fold cross validation was employed and accordingly 10% of each fold was always used for testing. When using a validation set, ¼ of the remaining data was used for validation and ¾ for training. Accordingly, a validation set would hold 22.5% and the training set 67.5% of the entire data set. In all experiments ¼of the training set was also used for early stopping, regardless of whether another validation set was used or not. The validation set was randomized before the training of each network, so each network was trained on slightly different data.In this study, the output from an ensemble was always the average of the output from all members. For all problems, a localist coding was used, so there was one output unit per class, and the unit with the highest (averaged) output determined the predicted class.

B. Ensemble setups

The six groups of ensemble setups used in the study were named 3-layered, 4-layered, Mixed, Selected, GA and All.In the 3-layered group, all ANNs had exactly one hidden layer. Three different setups were evaluated; a single

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 4, July 2014

ANN, five ANNs and ten ANNs. In this group all ANNs in an ensemble had identical topologies. The exact architecture was based on data set characteristics. More specifically the number of hidden units in the first layer was shown in **eqn. (1),**

$$h = \lfloor \sqrt{(v.c)} \rfloor \quad (1)$$

where v is the number of input variables and c is the number of classes. The 4-layered group was identical to the 3-layered, with the obvious exception that ANNs used here had two hidden layers. The number of hidden units is found below shown in **eqn. (2),**

$$h2 = \lfloor \frac{3}{2} h \rfloor \quad (2)$$

$$h2 = \lfloor \frac{3}{2} c \rfloor \quad (3)$$

where c again is the number of classes and h is calculated. The first two groups represent frequently used choices when identical networks are used to form the ensemble.In the mixed group, two different setups were evaluated. Here either five (M5) or ten (M10) ANNs, with randomized topologies, were combined to make up the ensemble. The randomized topology was based on the heuristics described above. Each ANN in the ensemble could have either one or two hidden layers. The number of hidden units in networks with one layerwas

$$m3h = \mathrm{h} + \lfloor rand.\mathrm{h} \rfloor \quad (4)$$

where M3 stands for Mixed3-layer, rand is a uniform random number in the range [0, 1] and h is, as before, calculated using **eqn.(1).** The numbers of hidden nodes in networks with two hidden layers were, for the first layer

$$m4h1 = \mathrm{h} + \lfloor rand.(\mathrm{h}/c) \rfloor \quad (5)$$

and for the second layer

$$m4h2 = \mathrm{h} + \lfloor rand.\left(\frac{\mathrm{h}}{c}\right) \rfloor + c \quad (6)$$

respectively. This group is potentially interesting since it represents a rather uncommon choice. The extra work needed, compared to the two previous groups is, however, quite marginal.In the selected group, 50 ANNs (25 with one hidden layer and 25 with two hidden layers) were trained. The exact architecture for each ANN was randomized according to the procedure described in **eqn. (4)-(6**). After training, the ANNs were sorted on validation set accuracy. The selected ensembles consisted of the single best (S1), the best five (S5) and the best ten (S10) ANNs. This setup is clearly more costly, since many more ANNs are trained. Although the approach is very straightforward, it appears to be extremely uncommon.

The last evaluated setup not based on GAs simply used all 50 trained ANNs in the ensemble. For a summary of the evaluated setups not using GAs, **see Table 2.**

Name	#ANNs	#Hidden layers	Identical topology	Selection
3-one	1	1	-	-
3-five	5	1	Yes	All
3-ten	10	1	Yes	All
4-one	1	2	-	-
4-five	5	2	Yes	All
4-ten	10	2	Yes	All
M5	5	1-2	No	All
M10	10	1-2	No	All
S1	1	1-2	-	Best from 50
S5	5	1-2	No	Best 5 from 50
S10	10	1-2	No	Best 10 from 50
All	50	1-2	No	All

Table 2: Properties for setups not using GAs.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 4, July 2014

The GA group, which is a novel approach to the creation of ensembles, contained six slightly different setups. Each setup used the same pool of 50 ANNs as the selected group. Here, however, the selection of members for the ensemble was based on GAs. More specifically, a possible combination of ANNs was represented as a chromosome with 50 genes, each gene denoting a specific ANN. Each chromosome was represented as a sequence of zeroes and ones (a bit string) where individual genes correspond to a specific ANN. As expected a "1" would indicate that the specific ANN should be included in the ensemble. All settings, except fitness function, were identical for all five setups. In each experiment stochastic uniform selection was used, the crossover and mutation rates were 0.8 and 0.05 respectively. The maximum number of generations was set to 200, but evolution was aborted if there was no change in fitness (for the most fit individual) over 30 consecutive generations. The populations consisted of 1000 individuals [5].

The first setup, called GA/Val, used accuracy on the validation set as fitness function. The second setup (GA/TrV) used accuracy on the training and validation sets as fitness function.

The third setup (GA/TaV) used accuracy on the training and validation sets as fitness function. During evolution the best individual (ensemble) from each generation is saved and after completion the "generation winner" with highest accuracy on the validation set was returned.

The fourth and fifth setups (GA/Tr3V and GA/Tr6V) are very similar to GA/TrV (**as shown in Table.3**). All three used the training and validation sets to calculate the fitness. In GA/TrV a correctly classified instance from the validation set was only worth exactly as much as a correct prediction on the training set. But in GA/Tr3V and GA/Tr6V a correctly classified validation set instance was weighted with a factor 3 or 6, respectively. The obvious motivation for this approach was to prioritize accuracy on data not used for training, while still using as much data as possible when calculating the fitness.

The sixth and final GA setup (GA/Test) used accuracy on the test set as fitness function. It must, of course, be noted that target values for a production set are by definition not available during construction of a model. Therefore this setup is not useful as a construction strategy but serves only as a demonstration of what level of accuracy a combination of the available ANNs could achieve. **Table.3** summarizes the evaluated setups using GAs.

Name	Fitness based on	Set used for selection
GA/Val	Validation	-
GA/TrV	Train/Validation	-
GA/TaV	Train/Validation	Validation
GA/Tr3V	Train/3*Validation	-
GA/Tr6V	Train/6*Validation	-
GA/Test	Test	-

Table 3: Properties for setups using GAs.

C. Results

The three tables below show the results from the experiments. Tabulated values represent average accuracy (on the test set) over all ten folds of each data set. Some of the 23 data sets used here were also used by Lim, Loh and Shih in [4]. The column named LLS lists the best results (**as shown in Table.4**) obtained by any algorithm in the LLS study, where altogether 33 different algorithms were evaluated. To calculate the normalized value (Norm.) the following procedure was used:

- For each single run the accuracy obtained by a specific setup was divided by the result obtained by GA/Test on the same run. This value presents the accuracy obtained as a percentage of "optimal" accuracy, represented by GA/Test.
- The calculated percentages from each run were averaged to produce a single, mean, value for each setup and data set.
- The values produced for each data set were again averaged to produce the single, tabulated, value for each setup.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 4, July 2014

The row average rank shows the average rank for the setup over all data sets and the row #Best shows the number of data sets where the setup produced the best result of all setups. It should be noted that ranks were calculated using more than the three tabulated decimals.

Data sets	3-one	3-five	3-ten	4-one	4-five	4-ten	LLS
BLD	0.703	0.683	0.700	0.661	0.692	0.686	0.72
Cleve	0.755	0.810	0.803	0.768	0.823	0.803	-
CMC	0.527	0.531	0.546	0.436	0.528	0.542	0.57
Crx	0.739	0.843	0.854	0.664	0.824	0.823	-
German	0.691	0.716	0.709	0.654	0.708	0.698	-
Glass	0.555	0.618	0.673	0.418	0.659	0.659	-
Hepatitis	0.777	0.812	0.818	0.794	0.806	0.806	-
Horse	0.745	0.824	0.840	0.734	0.784	0.795	-
Iono	0.854	0.941	0.924	0.724	0.908	0.941	-
Iris	0.920	0.940	0.960	0.947	0.947	0.960	-
Labor	0.800	0.850	0.850	0.775	0.825	0.838	-
Led7	0.639	0.732	0.736	0.612	0.732	0.735	0.73
Lymph	0.612	0.765	0.777	0.647	0.741	0.782	-
PID	0.738	0.749	0.763	0.710	0.748	0.753	0.78
Satellite	0.808	0.839	0.838	0.776	0.851	0.856	0.90
Sonar	0.746	0.814	0.832	0.746	0.786	0.823	-
TAE	0.418	0.494	0.553	0.441	0.494	0.529	0.67
Tic-Tac-Toe	0.719	0.744	0.780	0.816	0.797	0.792	-
Waveform	0.870	0.869	0.870	0.830	0.869	0.869	0.85
WBC	0.900	0.974	0.967	0.865	0.965	0.969	0.97
Vehicle	0.672	0.824	0.831	0.644	0.828	0.842	0.85
Wine	0.867	0.939	0.978	0.928	0.961	0.978	-
Zoo	0.673	0.909	0.936	0.718	0.927	0.900	-
Norm.	0.819	0.891	0.907	0.797	0.890	0.899	-
Average rank	15.26	10.78	7.50	16.13	11.47	9.54	
#Best	1	1	2	0	0	1	-

Table 4: Results for uniform ensembles

The most interesting observations from **Table 4** are:

- There was a large difference in accuracy between single ANNs and ensembles.
- An ensemble with only five members seems to be too small when using identical ANNs, although the difference was much less significant between the two ensembles than between the smaller ensemble and the single network.
- The best results for both Waveform and WBC were better than the best results achieved by LLS, while results for TAE were much worse. It should be noted, however, that ANNs in general always perform very poorly on the TAE data set, compared to, for instance decision trees.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 4, July 2014

Data sets	M5	M10	S1	S5	S10	All	LLS
BLD	0.667	0.708	0.700	0.711	0.711	0.706	0.72
Cleve	0.816	0.800	0.781	0.813	0.819	0.829	-
CMC	0.539	0.537	0.524	0.546	0.557	0.548	0.57
Crx	0.853	0.851	0.849	0.846	0.853	0.857	-
German	0.717	0.709	0.690	0.711	0.714	0.707	-
Glass	0.555	0.641	0.600	0.686	0.686	0.668	-
Hepatitis	0.818	0.812	0.829	0.853	0.824	0.818	-
Horse	0.818	0.816	0.816	0.845	0.845	0.853	-
Iono	0.916	0.949	0.914	0.938	0.938	0.935	-
Iris	0.940	0.947	0.940	0.947	0.967	0.960	-
Labor	0.800	0.838	0.838	0.863	0.863	0.850	-
Led7	0.726	0.736	0.731	0.733	0.734	0.732	0.73
Lymph	0.729	0.800	0.729	0.788	0.782	0.806	-
PID	0.733	0.756	0.752	0.754	0.761	0.758	0.78
Satellite	0.850	0.846	0.865	0.865	0.865	0.848	0.90
Sonar	0.800	0.809	0.791	0.805	0.818	0.827	-
TAE	0.447	0.488	0.465	0.524	0.512	0.471	0.67
Tic-Tac-Toe	0.774	0.772	0.833	0.861	0.862	0.805	-
Waveform	0.866	0.868	0.864	0.868	0.867	0.865	0.85
WBC	0.957	0.964	0.967	0.969	0.972	0.965	0.97
Vehicle	0.806	0.807	0.784	0.841	0.842	0.831	0.85
Wine	0.978	0.983	0.956	0.978	0.978	0.983	-
Zoo	0.855	0.873	0.909	0.918	0.946	0.927	-
Norm.	0.878	0.896	0.886	0.913	0.915	0.907	-
Average rank	12.58	9.87	12.59	6.93	5.57	7.60	
#Best	0	1	0	1	2	5	-

Table 5: Results for mixed and selected ensembles

There are several interesting results in **Table 5:**

- Ensembles built from selected networks were very accurate. S5 and S10overall performed remarkably well, obtaining high accuracy on most data sets.
- In this study the concept of using mixed architectures did not pay off. For example, both M5 and M10 (Shown in **Table. 5**) had worse accuracy than 3-ten.
- Using a very large ensemble (all) did produce rather high accuracy, although not quite as high as S5 and S10 (Shown in **Table. 5**). The large ensemble, however, achieved the highest accuracy overall on as many data sets as five.
- The results on the LLS data sets were overall rather good, again with the exception of TAE.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 4, July 2014

Data sets	Val	TrV	TaV	Tr3V	Tr6V	Test	LLS
BLD	0.703	0.686	0.703	0.708	0.720	0.811	0.72
Cleve	0.819	0.816	0.819	0.823	0.797	0.926	-
CMC	0.556	0.555	0.560	0.558	0.560	0.638	0.57
Crx	0.850	0.853	0.850	0.856	0.843	0.919	-
German	0.713	0.731	0.724	0.734	0.727	0.803	-
Glass	0.664	0.668	0.664	0.659	0.691	0.845	-
Hepatitis	0.824	0.841	0.847	0.841	0.847	0.888	-
Horse	0.832	0.824	0.821	0.824	0.824	0.895	-
Iono	0.941	0.943	0.943	0.941	0.951	0.973	-
Iris	0.967	0.967	0.967	0.960	0.973	1.000	-
Labor	0.888	0.863	0.875	0.888	0.875	0.938	-
Led7	0.736	0.737	0.736	0.737	0.736	0.762	0.73
Lymph	0.782	0.794	0.788	0.806	0.806	0.912	-
PID	0.751	0.765	0.760	0.756	0.748	0.830	0.78
Satellite	0.866	0.867	0.867	0.866	0.865	0.881	0.90
Sonar	0.827	0.832	0.841	0.836	0.832	0.936	-
TAE	0.524	0.512	0.535	0.518	0.541	0.694	0.67
Tic-Tac-Toe	0.868	0.879	0.879	0.873	0.878	0.920	-
Waveform	0.867	0.867	0.869	0.867	0.867	0.884	0.85
WBC	0.964	0.968	0.968	0.964	0.961	0.983	0.97
Vehicle	0.833	0.835	0.826	0.838	0.836	0.928	0.85
Wine	0.983	0.978	0.983	0.983	0.972	1.000	-
Zoo	0.927	0.909	0.946	0.927	0.936	1.000	-
Norm.	0.914	0.914	0.918	0.918	0.919	1.000	-
Average rank	6.56	5.70	4.48	4.91	5.65		
#Best	2	4	6	5	5	-	-

Table 6: Results for setups using GAs

The main results in **Table 6** are:

- Using the Norm. value, all GA setups had, at least, as high accuracy as any other setup evaluated. Looking at average ranks, all GA approaches outperformed all other setups, with the exception of S10, which had a lower average rank than Val.
- Leaving the TAE data set out, the results are comparable to the best results achieved in LLS.

To determine if there are statistically significant differences between the setups evaluated, a Friedman test (**Fig.1**) was performed. As seen in **Fig.1** below, this test showed no significant differences between any of the GA approaches, S5, S10, All, 3-ten, 4-ten and M10.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 4, July 2014

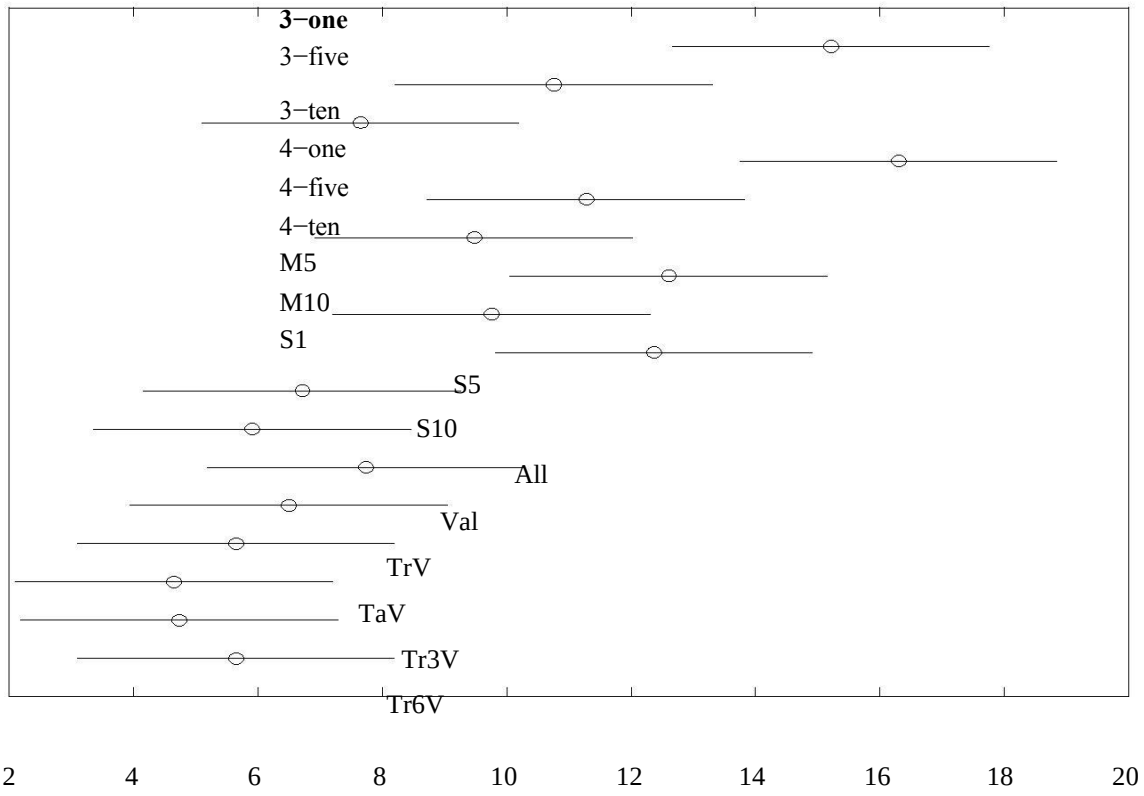

Fig 1: Friedman test Study 1

D. Results

The purpose of this study was to compare different ways of creating accurate ANN ensembles. The most important conclusion is that the option to somehow actively select the members of an ensemble appears to be a strong approach. The study showed that the simple procedure of selecting a fixed number of networks, based on validation set accuracy, from a large pool of ANNs, results in increased accuracy. For the practicing data miner this is a basic yet very effective approach.

The GA approaches did produce the most accurate ensembles, although the difference was quite small. On the other hand, the GA setups, especially when emphasizing the validation set, outperformed the non-GA setups on most data sets. It should be noted that the most straightforward GA-approach; i.e. to use only the validation set when calculating the fitness, was not very successful. The reason is probably that the GA is too powerful, leading to over fitting and poor generalization. The idea to actively search for the best members of an ensemble is very appealing. The GA-approach proposed here is easier to grasp and more intuitive than most similar methods, the reason being that the fitness is based directly on accuracy and applied to ensembles instead of single networks. In addition, it is an implicit advantage that ensembles of different sizes are continuously evaluated and compared The fact that the statistical test showed few significant differences is partly due to the procedure used; i.e. comparing so many approaches against each other. Having said that, the main picture, looking at the average ranks, is that the GA-approaches, S5, S10 (**shown in Table. 5**) and all were most successful. Somewhat surprising, the very simple method 3-ten also performed pretty well.

ISSN: 2319-5967
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume3, Issue 4, June 2014

III. CONCLUSION

A very interesting observation from the experimentation is that the ensemble selected by the Genetic Algorithm has much higher accuracy on the "fitness set" than any setup not using Genetic Algorithm. Unfortunately, this quite often failed to carry over to the test set.Most algorithms for ensemble creation do not explicitly target diversity. Instead different techniques are used to introduce implicit diversity. As described, implicit methods typically manipulate either the training data or some learning parameter. For ANN ensembles, some diversity is introduced just by normal randomization of the weights. In addition, ANN diversity is often produced by supplying each ANN with a slightly different training set, or by using different architectures for each ANN.

ACKNOWLEDGMENT

I would like to thank to all my lab assistants to provide me all the time lab facilities.

REFERENCES

[1] U. Johansson, T. Löfström and L. Niklasson, Obtaining Accurate Neural Network Ensembles, International Conference on Computational Intelligence for Modeling Control and Automation - CIMCA, Vienna,Austria, IEEE Computer Society, Vol. 2:103-108, 2005.

[2] U. Johansson, T. Löfström, R. König and L. Niklasson, Introducing GEMS – a Novel Technique for Ensemble Creation, 19th Florida Artificial Intelligence Research Society Conference (FLAIRS) 06, Melbourne Beach, FL, AAAI Press, pp. 700-705, 2006.

[3] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, Classificationand Regression Trees, Wadsworth International Group, 1984.

[4] T.-S.Lim, W,-Y.Loh and Y.-S Shih, A Comparison of Prediction Accuracy, Complexity, and Training Time and 33 Old and New Classification Algorithms, Machine Learning, 40:203-229, 2000.

[5]. Wikipedia - http://www.wikipedia.org,Ai-junkie - http://www.ai-junkie.com/

AUTHOR BIOGRAPHY

Tiruveedula Gopi Krishna, Received the B.Sc. and M.Sc.,M.E. M.Phil. Degrees in Computer Science Engineering from Andhra University, Anna University. Manav Bharati University (1997, 2001, 2004, 2010) respectively. Currently working as a computer faculty, Hoon Al-jufra, for Sirt University since 2007 to till date in Libya.Research area is Data Mining and Artificial Intelligence, Ph.D registered with Rayalaseema University, in 2009(Research Scholar) . He has got published 17 Research papers in reputed and Highest Impact Factor International Journals and 3 papers published as co-author in International Journals, and attended and published 4 research papers in an International conferences, and he has published a Book with entitled"Quick learn basics of Computer "recently.