CHAPTER 6

PERFORMANCE ANALYSIS

Assessment is necessary to know the correctness of any system and this is done through this chapter which evaluates the performance of the emotion recognition system and performs a comparative study.

6.1 PERFORMANCE EVALUATION

Performance for classification tasks can be defined in terms of the correct rate of classification for a set of test cases. The performance standard for comparing a computer-based system has two views:

- Based on human expert performance.
- Based on best existing system.

The former view is adapted for grading the performance of the candidate system as other existing systems with the same methodology couldn't be traced.

In a classification task, the precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class). A precision score of 1.0 for a class C means that every item labeled as belonging to class C does indeed belong to class C.

	Actual Class (expectation)	
Predicted class (observation)	**t_p** (true positive) Correct result	**f_p** (false positive) Unexpected result
	f_n (false negative) Missing result	**t_n** (true negative) Correct absence of result

Table 6.1: Performance Evaluation Measures

In classification tasks, the terms true positives, true negatives, false positives, and false negatives as stated in Table 6.1 compare the results of the classifier i.e., Neuro-Fuzzy System under test with trusted external judgments (feedback of ANNOTATOR). The terms *positive* and *negative* refer to the classifier's prediction and the terms *true* and *false* refer to whether that prediction corresponds to the external judgment.

The **precision** of a measurement system, also called reproducibility is the degree to which repeated measurements under unchanged conditions show the same results.

$$\text{Precision} = (t_p \ / \ t_p + f_p)$$

Other related measure used in classification is **accuracy**. Accuracy is also used as a statistical measure of how well a binary classification test correctly identifies or excludes a condition. The accuracy is the

proportion of true results (both true positives and true negatives) in the population. It is a parameter of the test.

$$\text{Accuracy} = (t_p + t_n) \,/\, (t_p + t_n + f_p + f_n)$$

The Predicted Class (observation) here refers to the Neuro-Fuzzy System and the Actual Class (expectation) refers to the Annotator.

SAMPLE 1: For a sample of 39 events (see Appendix B) which consists all the six emotions, neutral and a combination of multiple emotions the values are as follows :

$t_p = 32$, $f_p = 5$, $f_n = 2$, $t_n = 0$. The values indicate that the system has generated results for 37 events ($t_p = 32$ and $f_p = 5$) and failed to generate a result for two events ($f_n = 2$ and $t_n = 0$). Out of the results, those of 32 events have matched the feedback of Inter-Annotator-Agreement ($t_p = 32$) and those of five events do not match the Inter-Annotator-Agreement ($f_p = 5$).

Precision = $t_p / (t_p + f_p) = 32 / (32 + 37)$ = **0.86**

Accuracy = $(t_p + t_n) / (t_p + t_n + f_p + f_n) = (32 + 0)/(32+0+5+2)$

= **0.82**

SAMPLE 2: For a sample of 101 events with all the six emotions, neutral and a combination of multiple emotions, the values are as follows :

$t_p = 71$, $f_p = 16$, $f_n = 11$, $t_n = 3$. The values indicate that the system has generated results for 87 events ($t_p = 71$ and $f_p = 16$) and failed to generate a result for fourteen events ($f_n = 11$ and $t_n = 3$). Out of the results, those

of 71 events have matched the feedback of Inter-Annotator-Agreement (t_p = 71) and those of sixteen events do not match the Inter-Annotator-Agreement (f_p = 16). For three events neither the system nor the annotators were able to identify the emotion.

Precision = t_p / (t_p + f_p) = 71 / (71 + 16) = **0.82**

Accuracy = (t_p + t_n) /(t_p + t_n + f_p + f_n) = (71 + 3)/(71+3+16+11)

= **0.73**

Attribute Agreement Analysis is used to assess the agreement between the ratings made by appraisers (annotators) and the known standards (system). Attribute Agreement Analysis determines the accuracy of the assessments made by appraisers and identifies which items have the highest misclassification rates. This is performed using the statistical software used for quality improvement - **MINITAB 16** (from MINITAB Inc., USA)**.** If the ratings of appraisers and system agree, the possibility exists that the ratings are accurate. If there is disagreement, the rating usefulness is limited.

Agreement between readers is quantified by the Kappa (K) statistic:

- ➔ K is 1 when there is perfect agreement between the classification system
- ➔ K is 0 when there is no agreement better than chance

➔ K is negative when agreement is worse than chance

Value of K	**Strength of agreement**
< 0.20	Poor
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Good
0.81 - 1.00	Very good

For a sample of 39 events (see Appendix B) which is a collection of events with single, multiple and neutral emotions and different intensities and polarities, the results generated by the system and the annotator are tabulated as follows :

EVENT	APPRAISER	EMOTION
1	SYSTEM	HAPPINESS
1	ANNOTATOR	HAPPINESS
2	SYSTEM	DISGUST
2	ANNOTATOR	DISGUST
3	SYSTEM	SURPRISE
3	ANNOTATOR	SURPRISE
4	SYSTEM	DISGUST
4	ANNOTATOR	DISGUST
5	SYSTEM	NEUTRAL
5	ANNOTATOR	HAPPINESS
6	SYSTEM	HAPPINESS
6	ANNOTATOR	HAPPINESS
7	SYSTEM	DISGUST
7	ANNOTATOR	DESPAIR
8	SYSTEM	SURPRISE
8	ANNOTATOR	SURPRISE
9	SYSTEM	FEAR
9	ANNOTATOR	FEAR
10	SYSTEM	ANGER
10	ANNOTATOR	ANGER
11	SYSTEM	DESPAIR
11	ANNOTATOR	DESPAIR
12	SYSTEM	ANGER
12	ANNOTATOR	ANGER

EVENT	APPRAISER	EMOTION
13	SYSTEM	ANGER
13	ANNOTATOR	ANGER
14	SYSTEM	FEAR
14	ANNOTATOR	DISGUST
15	SYSTEM	HAPPINESS
15	ANNOTATOR	HAPPINESS
16	SYSTEM	DESPAIR
16	ANNOTATOR	HAPPINESS
17	SYSTEM	NEUTRAL
17	ANNOTATOR	NEUTRAL
18	SYSTEM	HAPPINESS
18	ANNOTATOR	DESPAIR
19	SYSTEM	DESPAIR
19	ANNOTATOR	DESPAIR
20	ANNOTATOR	HAPPINESS
20	ANNOTATOR	HAPPINESS
21	SYSTEM	ANGER
21	ANNOTATOR	ANGER
22	SYSTEM	DESPAIR
22	ANNOTATOR	DESPAIR
23	SYSTEM	ANGER
23	ANNOTATOR	ANGER
24	SYSTEM	DESPAIR
24	ANNOTATOR	DESPAIR
25	SYSTEM	ANGER
25	ANNOTATOR	ANGER
26	SYSTEM	DESPAIR
26	ANNOTATOR	DESPAIR
27	SYSTEM	SURPRISE
27	ANNOTATOR	SURPRISE
28	SYSTEM	FEAR
28	ANNOTATOR	FEAR
29	SYSTEM	SURPRISE
29	ANNOTATOR	SURPRISE
30	SYSTEM	FEAR
30	ANNOTATOR	FEAR
31	SYSTEM	ANGER
31	ANNOTATOR	DISGUST
32	SYSTEM	NEUTRAL
32	ANNOTATOR	HAPPINESS
33	SYSTEM	FEAR
33	ANNOTATOR	FEAR
34	SYSTEM	HAPPINESS
34	ANNOTATOR	HAPPINESS

EVENT	APPRAISER	EMOTION
35	SYSTEM	DISGUST
35	ANNOTATOR	DISGUST
36	SYSTEM	DESPAIR
36	ANNOTATOR	DESPAIR
37	SYSTEM	SURPRISE
37	ANNOTATOR	SURPRISE
38	SYSTEM	DISGUST
38	ANNOTATOR	DISGUST
39	SYSTEM	FEAR
39	ANNOTATOR	FEAR

Attribute Agreement Analysis for EMOTION

Date of study: 03-03-2012
Reported by: G.SHARADA
Name of product: EMOTION

Between Appraisers

Assessment Agreement

# Inspected	# Matched	Percent	95% CI
39	32	82.05	(66.47, 92.46)

Matched: All appraisers' assessments agree with each other.

Fleiss' Kappa Statistics

Response	Kappa	SE Kappa	Z	P(vs > 0)
ANGER	0.90769	0.160128	5.6685	0.0000
DESPAIR	0.75238	0.160128	4.6986	0.0000
DISGUST	0.68250	0.160128	4.2622	0.0000
FEAR	0.89417	0.160128	5.5841	0.0000
HAPPINESS	0.65179	0.160128	4.0704	0.0000
NEUTRAL	0.47297	0.160128	2.9537	0.0016
SURPRISE	1.00000	0.160128	6.2450	0.0000
Overall	0.78738	0.068220	11.5419	0.0000

Cohen's Kappa Statistics

Response	Kappa	SE Kappa	Z	P(vs > 0)
ANGER	0.90780	0.159446	5.6935	0.0000
DESPAIR	0.75264	0.159583	4.7163	0.0000
DISGUST	0.68293	0.159231	4.2889	0.0000
FEAR	0.89431	0.159231	5.6164	0.0000
HAPPINESS	0.65333	0.157704	4.1428	0.0000
NEUTRAL	0.48000	0.136776	3.5094	0.0002
SURPRISE	1.00000	0.160128	6.2450	0.0000
Overall	0.78788	0.067514	11.6698	0.0000

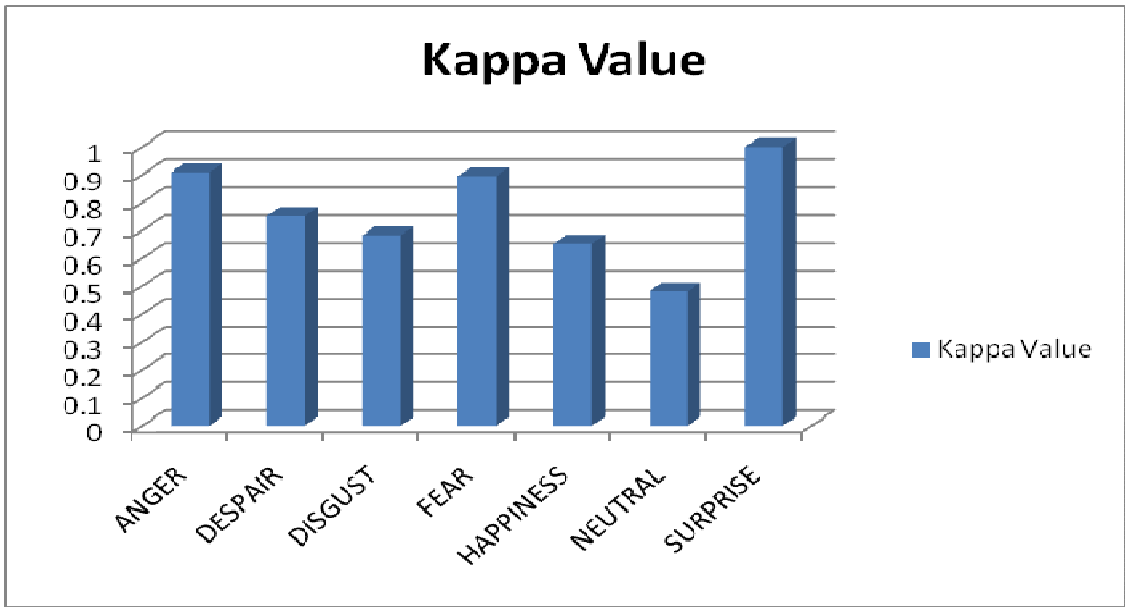

Fig 6.1: Graph for Kappa Statistic taken over 39 events

From the above analysis it can be seen that the correlation between the system and the inter-annotator agreement ranges from 0.48 for NEUTRAL (moderate) to 1.00 for SURPRISE (very good).

For a sample of 101 events, results were generated for 71 events and for 14 events the result was missing. Hence analysis can be done only for events with some result and is given as follows :

Attribute Agreement Analysis for EMOTION

```
Date of study:    15-03-2012
Reported by:      G.SHARADA
Name of product:  Emotion Recognition
Misc:
```

Between Appraisers

```
Assessment Agreement

# Inspected  # Matched  Percent      95% CI
         87         71    81.61  (71.86, 89.11)

# Matched: All appraisers' assessments agree with each other.


Fleiss' Kappa Statistics

Response      Kappa  SE Kappa        Z  P(vs > 0)
```

ANGER	0.859866	0.107211	8.0203	0.0000
DESPAIR	0.738233	0.107211	6.8858	0.0000
DISGUST	0.710345	0.107211	6.6257	0.0000
FEAR	0.895933	0.107211	8.3567	0.0000
HAPPINESS	0.758333	0.107211	7.0733	0.0000
NEUTRAL	0.213855	0.107211	1.9947	0.0230
SURPRISE	0.956160	0.107211	8.9185	0.0000
Overall	0.782023	0.045804	17.0733	0.0000

Cohen's Kappa Statistics

Response	Kappa	SE Kappa	Z	P(vs > 0)
ANGER	0.859903	0.107094	8.0294	0.0000
DESPAIR	0.738739	0.106537	6.9341	0.0000
DISGUST	0.710963	0.106386	6.6829	0.0000
FEAR	0.896057	0.106631	8.4034	0.0000
HAPPINESS	0.759225	0.105821	7.1746	0.0000
NEUTRAL	0.234604	0.068997	3.4002	0.0003
SURPRISE	0.956171	0.107108	8.9272	0.0000
Overall	0.782670	0.045185	17.3213	0.0000

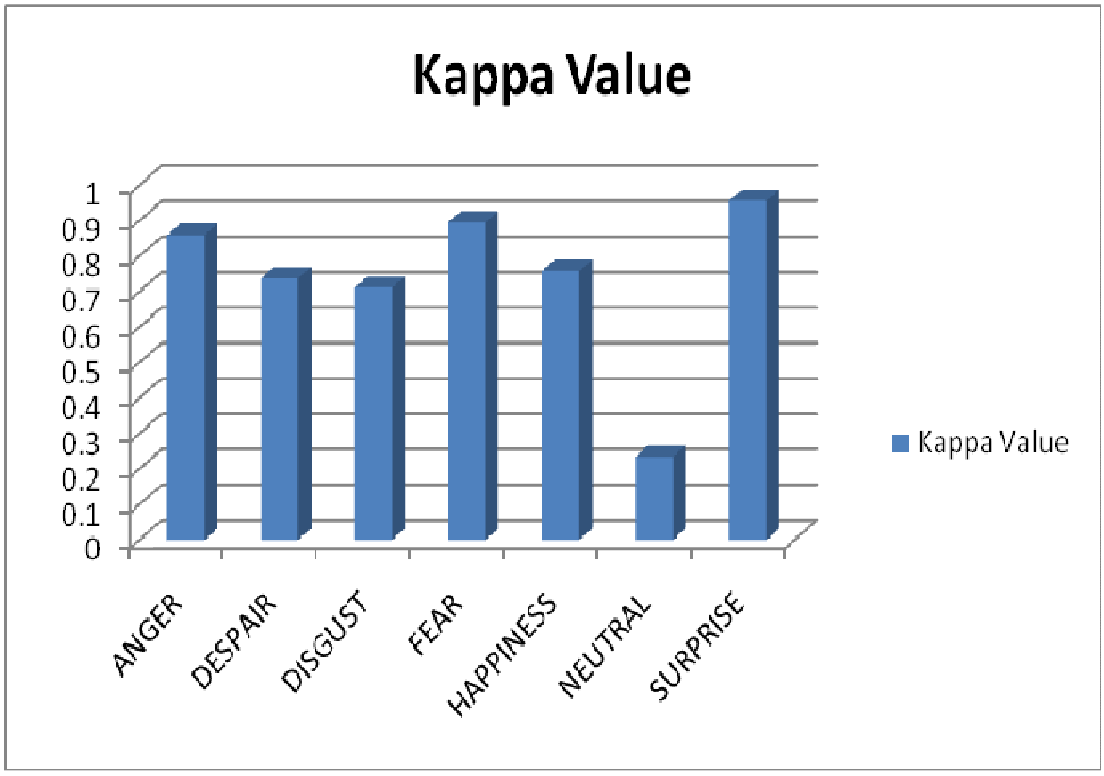

Fig 6.2: Graph for Kappa Statistic taken over 87 events

In this case it is observed that the correlation is minimum for the NEUTRAL emotion (0.23 –> fair) and maximum for the emotion SURPRISE (0.95 –> very good).

6.2 COMPARITIVE STUDY

ISEAR (International Survey on Emotion Antecedents and Reactions) project, was directed by Klaus R. Scherer and Harald Wallbott. Student respondents, both psychologists and non-psychologists, were asked to report situations in which they had experienced all of 7 major emotions (joy, fear, anger, sadness, disgust, shame, and guilt). The final data set thus contained reports on seven emotions each by close to 3000 respondents in 37 countries on all 5 continents.

This dataset [Appendix-C] was used by researchers in the domain of emotion recognition and was found to be the potential candidate for this study, in comparison to the other emotion recognition systems. Some of the statements related to the five basic emotions of the proposed system were taken randomly from the dataset and subjected to the emotion recognition system. The outputs were generated for the sentences which had an emotional word in it. But for sentences without an emotional word, the output was NEUTRAL. The comparative study is indicated by the following table:

EMOTION	NO.OF INSTANCES	RECOGNITION ACCURACY
HAPPINESS	50	70
DESPAIR	50	76
ANGER	40	80
FEAR	40	82
DISGUST	40	78

Table 6.2: Emotion recognition accuracy for ISEAR dataset sample

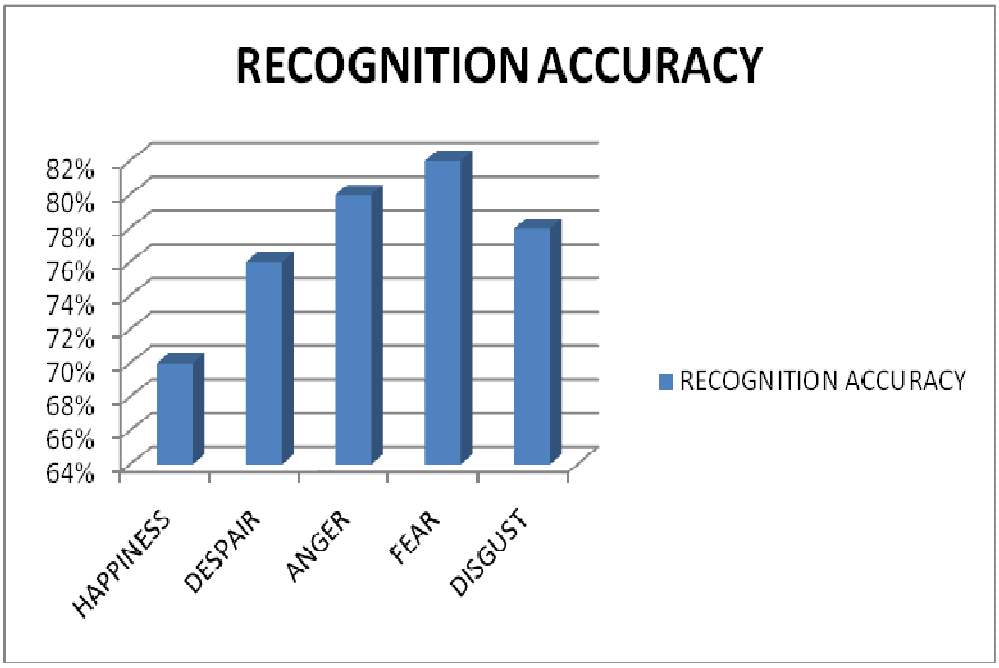

Fig 6.3: Graph for emotion recognition accuracy for ISEAR sample

The recognition accuracy for the emotion “Fear” was found to be more than the other emotions as the sentences had more probability of occurrences for the emotion keywords expressing fear. The emotion of “Happiness” gave the minimum accuracy as it was expressed implicitly in most of the sentences in the dataset and not explicitly through emotional keywords.