KNOWLEDGE BASED SELECTION IN SOFT COMPUTING MODELS USING PARETO CHART AND PARETO PRINCIPLE

Anil Kumar S

Assistant, EE II S, University of Kerala, Thiruvananthapuram, Kerala, India

Abstract

Pareto Chart is a statistical tool which is helpful for studying about exceptional data groups based on the variation in their frequency values. Analysis of data is done based on combining the models of frequency distribution and cumulative frequency distribution. Being a major tool in Quality Management, Pareto chart has a wide range of applications in project management and project engineering. Also, in Soft Computing Projects, its concept can be utilized for deciding among several possible alternatives based on prior knowledge. This tool can have a significant use in project situations which demand stochastic decision making. This article illustrates the empirical use of concepts derived from Pareto chart and Pareto principle specifically in Soft Computing Models, for making balanced decisions based on statistical knowledge acquired through prior experience.

Keywords: *Pareto Chart, Pareto Principle, Statistical Analysis, Stochastic Decision Making, Soft Computing.*

---***---

1. INTRODUCTION

In Soft Computing Applications, as in real life, there are situations which demand for making decisions with incomplete knowledge and/or limited resources. In such situations, the selection of the most fruitful option among several available alternatives will become more challenging. When the decision makers are not able to logically connect the available knowledge with their selection policy for upcoming events, the only possibility may be guessing. The selections among different options done through such blind guesses are realized in automated system simply by generating random numbers as being done in most of the conventional gaming systems..

However, most of the practical situations demand for more intelligent approaches in making such selections. Rather than making wild guesses in a blind manner through simple random selection, intelligent guesses are preferable when aiming to improve the ultimate outcome. Such guesses based on prior experience are known as Educated Guesses or Learned Guesses. In learned guesses, the selection process utilizes prior knowledge and makes more intelligent decisions which help to bring more beneficial results.

Further, while making such choices in a stochastic manner, most of the soft computing applications need not select the best option in all occasions. It is being done so, primarily for experiencing sub-optimal and non-optimal alternatives also for the purpose of adaptive learning. However in critical situations, the intelligent systems have to give more priority in selecting the most optimal options available at that moment.

Selecting or rejecting the extreme options can be done quite easily by sorting the options based on their score values derived from prior events. This can be implemented by simple query or a simple loop in the program. But the adaptive learning systems demand to occasionally select every possible option; but proportionate to the score values assigned based on prior experience. Thus the system insists to to give more chance to more beneficial options and at least some chance for even the least desirable ones. So the selection becomes a stochastic process.

Usually this selection is done in a probabilistic manner based on the analysis of prior outcomes generated by such events. The statistical analysis of probable options by using their probabilities and the selection process by using a stochastic value can be illustrated easily with a Pareto chart and Pareto principle. The following sections illustrate the materials and methodologies used for this experimental study.

2. ELEMENTARY CONCEPTS

Some of the elementary concepts which are used for deriving solution models in this study are mentioned in this section.

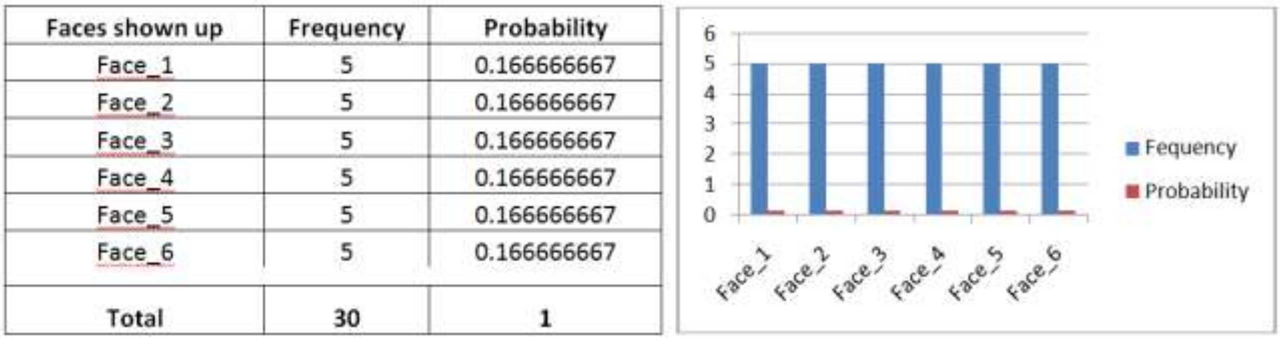
2.1 Stochastic Processes

Stochastic process indicates the one whose outcomes can be statistically analyzed but cannot be predicted. Most of the statistical problems in real world fall under this category.

2.2 Frequency Distribution

Frequency distribution gives a clear idea on number of occurrences of the events. Frequency distribution helps to analyze the events based on their repeating nature during the previous experience.

If the counts of occurrence for all the events are equal, the graphical representation shows the bars at same level. Same can be seen for a probability distribution of equally likely events. Line graph will show a straight line parallel to the x-axis. This is illustrated in Figure-1 which represents the frequency distribution of throwing a fair die.

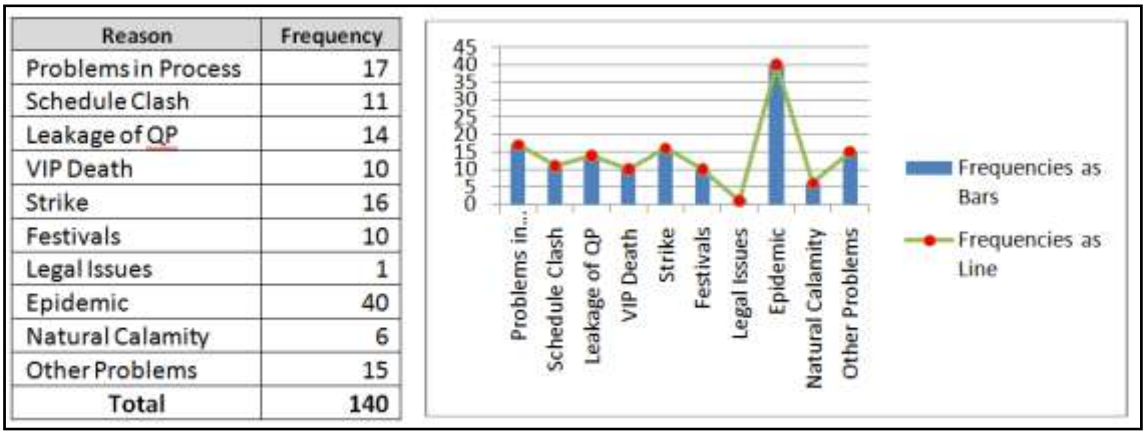


Faces shown up	Frequency	Probability
Face_1	5	0.166666667
Face_2	5	0.166666667
Face_3	5	0.166666667
Face_4	5	0.166666667
Face_5	5	0.166666667
Face_6	5	0.166666667
Total	30	1

Fig -1: Frequency distribution of Equally Likely Events as in throwing a fair die.

However, in most of the practical cases in real life, the situation is different. The equally likely events are very rare. In such cases, the frequency distribution will show the up and downs and it becomes more challenging to derive analytical results.

Consider the following example of analyzing the data regarding the possible reasons to postpone examinations after fixing a schedule. (Disclaimer: Randomly generated Sample data is used for this study so as to avoid legal consequences of using real data having confidential nature for the purpose of academic research and learning).

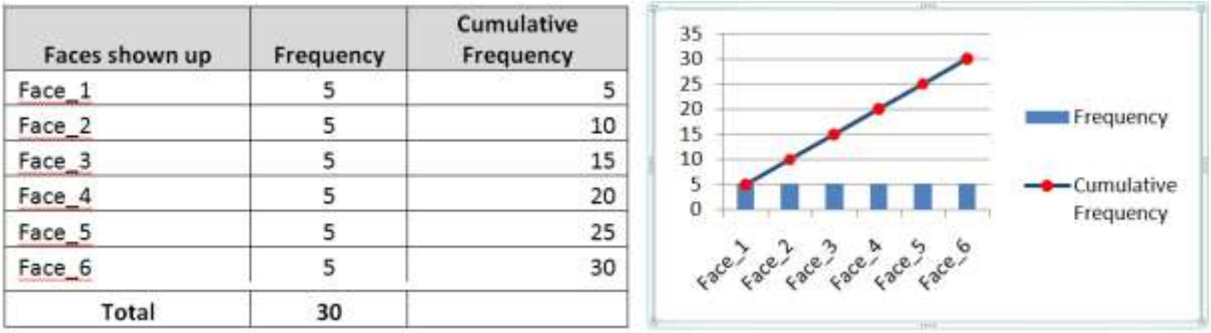


Reason	Frequency
Problems in Process	17
Schedule Clash	11
Leakage of QP	14
VIP Death	10
Strike	16
Festivals	10
Legal Issues	1
Epidemic	40
Natural Calamity	6
Other Problems	15
Total	140

Fig -2: Frequency Distribution of events with unequal probabilities

3. Cumulative Frequency Distribution

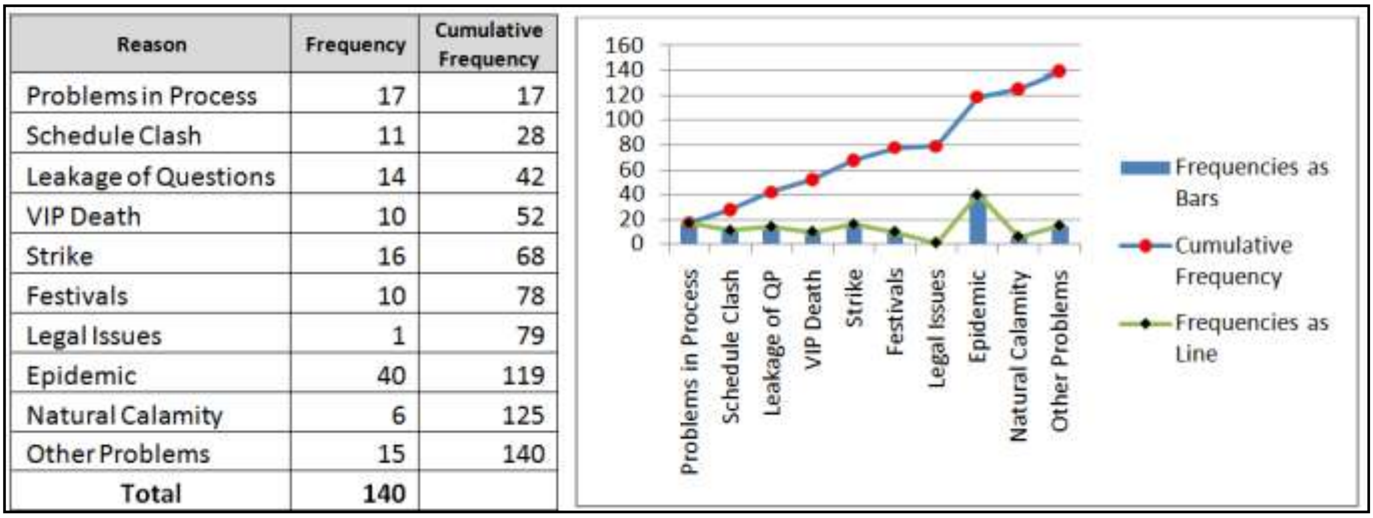
Cumulative Frequency represents the progressive totals of the Frequencies. In the case of events with equal frequency values, as in throwing a fair die, the cumulative frequencies form a straight line in the graph.



Faces shown up	Frequency	Cumulative Frequency
Face_1	5	5
Face_2	5	10
Face_3	5	15
Face_4	5	20
Face_5	5	25
Face_6	5	30
Total	30	

Fig -3: Cumulative Frequency Distribution of Equally Likely events, as in throwing of a fair die

Since most of the practical situations deal with events having unequal frequency values, the cumulative frequency will not show a straight line in the graph. It will be a sequence of line segments of different slopes. The position under each of these line segments represents the slot corresponding to the particular event related to the range of cumulative frequency values.



Reason	Frequency	Cumulative Frequency
Problems in Process	17	17
Schedule Clash	11	28
Leakage of Questions	14	42
VIP Death	10	52
Strike	16	68
Festivals	10	78
Legal Issues	1	79
Epidemic	40	119
Natural Calamity	6	125
Other Problems	15	140
Total	140	

Fig -4: Cumulative Frequency Distribution of events with unequal probabilities

The major deviations in the slope of line segments in the cumulative data line indicates major deviation in the frequency value of the particular event with respect to the frequencies of neighboring events and the general trend of the graph.

Higher the slope of the line segment, higher will be the number of frequency contributed by the related event.

Pareto chart is a combination of the bar chart representing the frequencies and the line graph representing the cumulative frequencies. This tool is named after ‘Vilfredo Pareto’ a famous Italian scientist.

In its popular form, the events are sorted in the reverse order of frequency values. The one with highest frequency value is shown in the left most ends.

3.1 Pareto Chart

Table -1: Data sorted in the decreasing order of frequency values

SL No	Reason	Frequency	Cumulative Frequency	Cumulative Percentage
1	Epidemic	40	40	28.57142857
2	Problems in Process	17	57	40.71428571
3	Strike	16	73	52.14285714
4	Other Problems	15	88	62.85714286
5	Leakage of Questions	14	102	72.85714286
6	Schedule Clash	11	113	80.71428571
7	VIP Death	10	123	87.85714286
8	Festivals	10	133	95
9	Natural Calamity	6	139	99.28571429
10	Legal Issues	1	140	100
Total		**140**		

The simplest model of Pareto chart is represented in Chart-1. As the data elements are sorted according to the decreasing order of frequency, the cumulative frequency will get presented as in the form a curve.

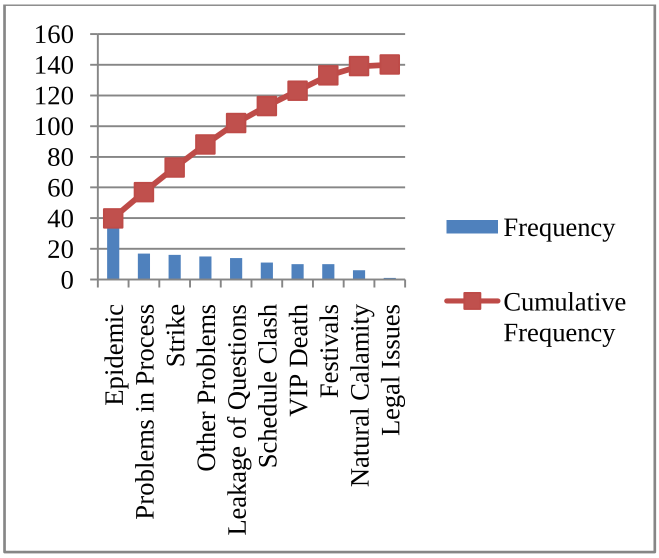


Chart -1: Basic Model of Pareto Chart

The above basic model of Pareto chart can be used as meta-model for deriving the advanced models by enhancing and modifying its features for solving various problems which involves comparative study of frequencies and utilizing such features for deriving solutions.

A model which is used for implementing Learned Guess in Soft Computing Applications is illustrated in the subsequent sections.

3.2 Pareto Principle

Pareto principle is a funny model of concept of Pareto chart, especially with the logic of cumulative frequency for data analysis. It is very popular in the field of Management Studies with the name 80:20 Rule.
Some of the popular versions of this 80:20 rule are listed below.

- 80% of problems are caused by 20% of persons and 20% of problems are caused by the remaining 80% persons.
- 80% of work in a team is done by 20% of members and only 20% of the work is done by remaining 80% of members.
- 80% of the profit is resulted from 20% of the products and 20% of the profit is resulted from 80% of the products.

Similar versions are heard in several contexts mainly to illustrate helpless and pathetic conditions in an organization or team.

Though seems to be a little funny and exaggerated, this concept makes sense in presenting the actual picture of a scenario in general. The figures 80 and 20 can vary, but it is an observable fact that majority of the resources make minor portion of the outcomes and vice-versa in several practical situations of real life too.

4. PARETO CHART FOR PARETO PRINCIPLE

In this context, it is quite interesting to illustrate the Pareto Principle using a Pareto chart. This will give a special outlook on the empirical use of Pareto charts in decision making situations.

For realizing the benefits of ‘Pareto concept’, the following illustration uses the problem of analyzing the reason for postponement of examinations, but with data values suitable to express the Pareto Principle. Here the reasons are classified into two groups, internal reasons and external reasons.

20% of the reasons are taken as internal reasons and given 80% of the total frequency values.
So re-sampling the data is done as follows.

Table -2: Data grouped based on nature of reasons (internal and external)

SL No	Reason	Frequency	Cumulative Frequency	Nature of Reason
1	Problems in Process	50	50	Internal Reasons
2	Schedule Clash	30	80	
3	Strike	3	83	External Reasons
4	Legal Issues	2	85	
5	Leakage of Questions	4	89	
6	Epidemic	1	90	
7	VIP Death	2	92	
8	Festivals	3	95	
9	Natural Calamity	4	99	
10	Other Problems	1	100	
Total		**100**		

The Pareto chart of the above data is given in Chart-2

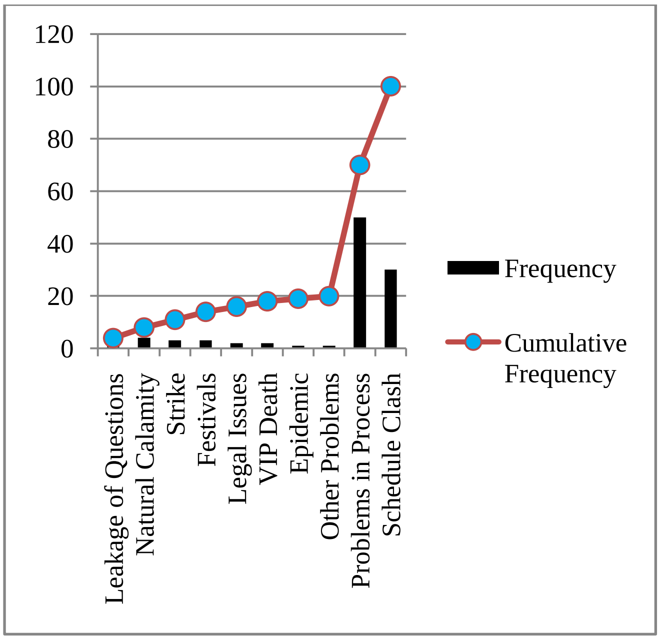


Chart -2: Pareto chart for Pareto Principle

The 80:20 concept present in the above Pareto Chart is illustrated in the Figure-5. Here the 20% of y-values are distributed against 80% of the x-values and 80% of the y-values are distributed against 20% of the x-values.

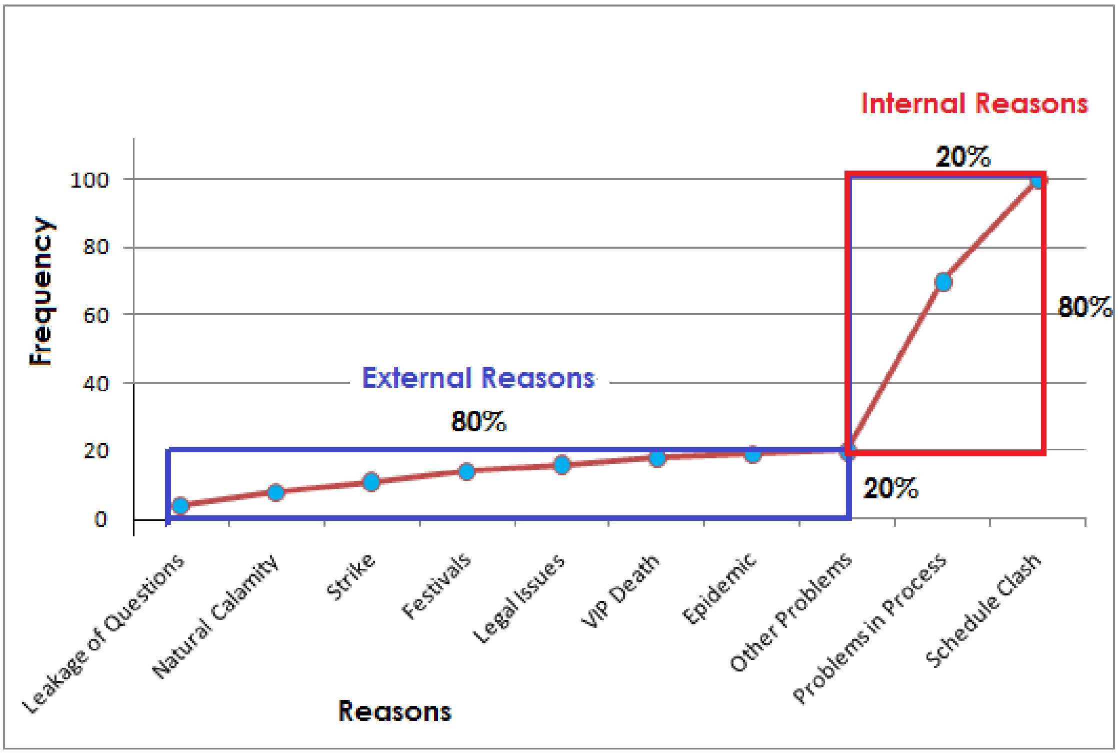


Fig -5: The Pareto Principle visualized through Pareto Chart.

Here the Internal reasons (20% among the entire set of reasons) occurs during 80% of the total count. It indicates that the organization can apply some remedial measures on those 20% reasons which can lead to reduce the occurrences of 80% problems.

On the other hand, 80% of the reasons depend on external factors, on which the organization has very less control over them. However, as they are less (20%) in number, there is not much harm in swimming according to their flow, i.e. nothing other than postponing the examination.

If the 80:20 ratio become 20:80, the scenario will become more funny because the organization and the society reaches to a pathetic situation as shown in the cartoon in Fig-6.

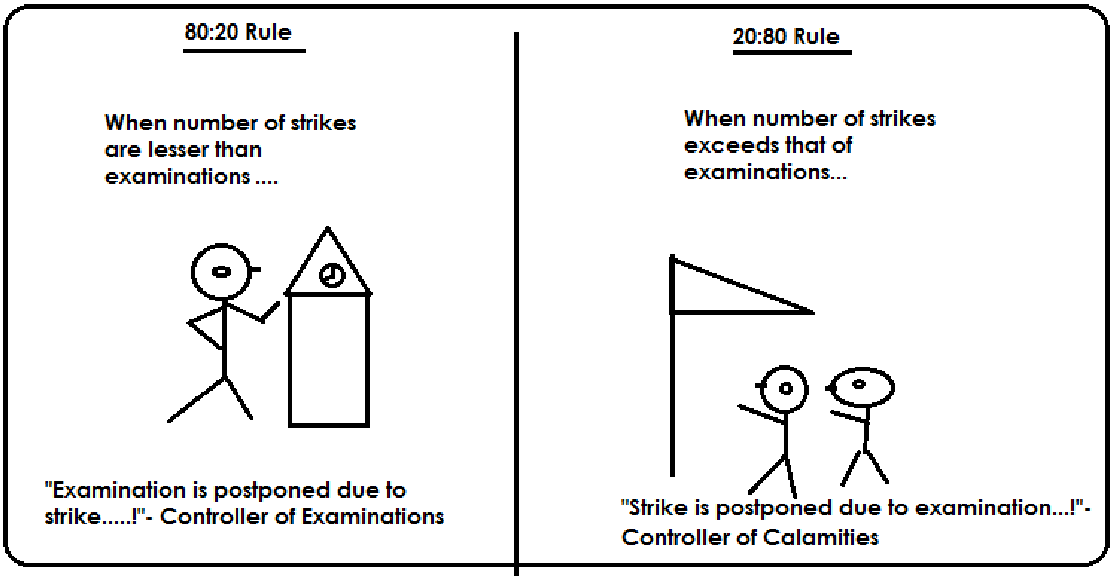

80:20 Rule
When number of strikes are lesser than examinations ....
"Examination is postponed due to strike.....!"- Controller of Examinations
20:80 Rule
When number of strikes exceeds that of examinations...
"Strike is postponed due to examination...!"- Controller of Calamities

Fig -6: Cartoon for 80:20 vs. 20:80.

In this cartoon the first half is the normal scenario, where the Controller of examinations enjoys the power to prepare Examination Calendars and reschedule the examination on unavoidable circumstances. In the second situation, it is the opposite extremity because of the number of strike goes beyond the number of examinations. Then the Examination controller will have nothing to do other than swimming with the flow or making some big decisions. Then the Controllers of Strikes and other calamities have to show the kindness to postpone or reschedule their strikes when examinations come..!

5. STOCHASTIC SELECTION PROCESS

The aforesaid peculiarities of Pareto chart indicate its scope for visualizing the stochastic selection among the available alternatives proportional to their prior occurrences and the values set for controlling parameters.

The decision making situations in Machine Learning Projects are illustrated using Markov Models. At each state decision is to be taken regarding the selection of the most suitable path leading to the next state. The available paths from a state represents the available choices, each of them are earmarked with the probability of getting selected based on available knowledge.

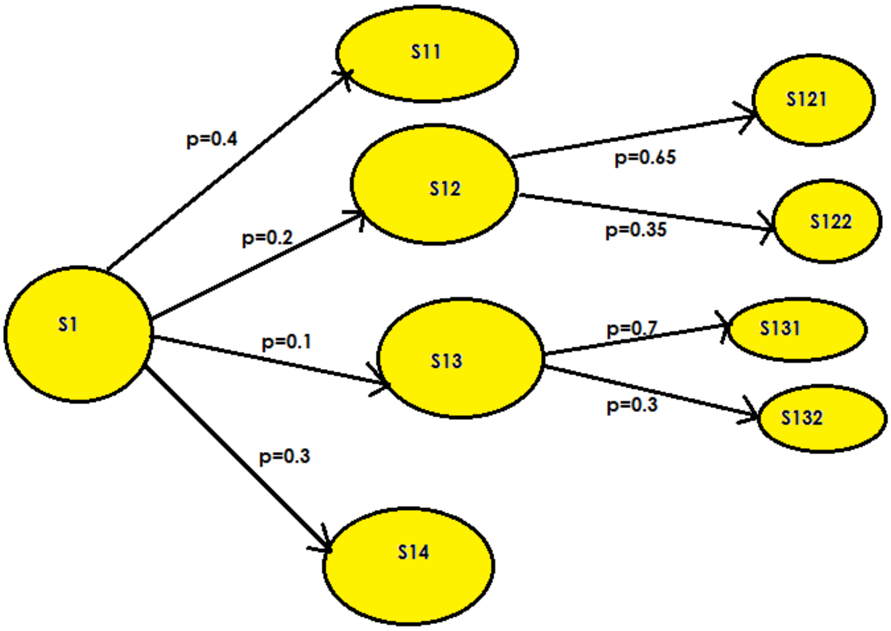


Fig -7: Example Markov Decision Process Model

The probability of choices available at the State S1 is tabulated below.

Table -3: The Probabilistic Score of 4 different options available at State S1

SL No	Neighboring State to be Selected	Probability	Probabilistic Score	Cumulative Score
1	S11	0.4	40	40
2	S12	0.2	20	60
3	S13	0.1	10	70
4	S14	0.3	30	100

The above data can be represented using a Pareto Chart as follows.

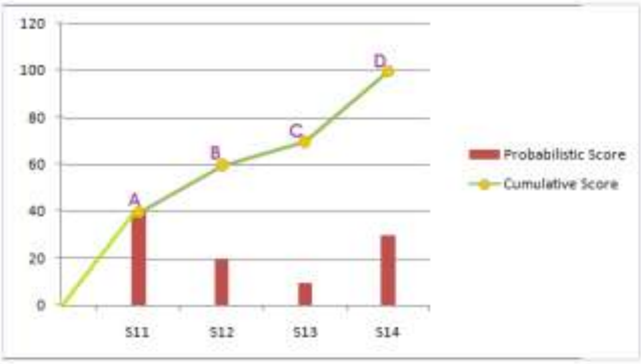


Chart -3: Pareto chart for the probabilistic selection

From the illustration in Chart-3, it can be seen that the cumulative score values formulate the sequence of line segments representing each of the four options along the x-axis and the length proportionate to their probabilistic score along the y-axis.

So counts of y-intercepts of points on the Line segments OA_AB_BC_CD are in the proportion 40_20_10_30.

For making a selection through stochastic decision process, a random point is to be taken in the range of values in y-axis. Here this range is 0 to 100. A random number selected in the range [0,100] can be the Stochastic Decision Pointer and is denoted by S. It is possible to find its position in the y-axis and find the slot on x-axis related to the corresponding line segment.

The aim is to choose one among the four options marked in the x-axis , but with a probability proportional to its score. For implementing this, the Pareto Chart can be used as a design model. So a random point is to be found out in the y-axis and find the corresponding slot in the x-axis. So the y-axis also has to be divided into four slots, proportionate to the probabilistic score.

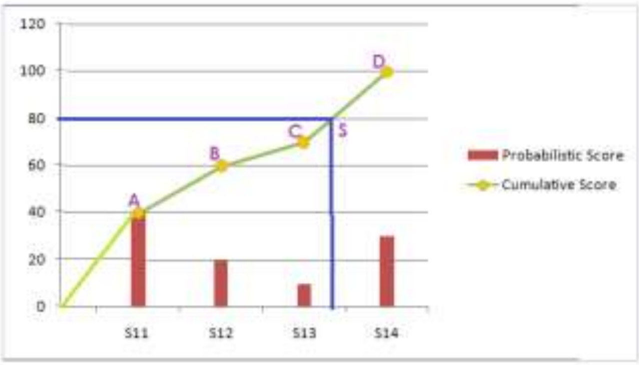


Chart -4: Selection through Stochastic Decision Pointer

Let the Stochastic Decision Pointer is 80. This value on the y-axis is related to the line segment CD, which represents the fourth slot in the x-axis i.e. that of the fourth option S14. This selection process can be tabulated as follows.

Table -4: The Probabilistic Score of 4 different options available at State S1

SlotNo	Range of values in y-axis	Line Segment in Pareto chart	Related Slot in x-axis	Path being selected
1	00..40	OA	1	S11
2	40..60	AB	2	S12
3	60..70	BC	3	S13
4	70..100	CD	4	S14

The above selection process with 4 options can be simulated by scaling to any number of options.

The above selection process can be used also for making Learned Guesses in the decision making situations in Soft Computing Projects based on the probabilities of options derived from knowledge earned through prior experience.

6. EXPERIMENTS IN GAMING

Stochastic Decision Making process explained above has been successfully tested in a chess gaming meta-model based on Reinforcement Learning. The prior knowledge is stored in the form of Q-values which represent the logical nearness towards a successful goal state. The negative Q-values are used for representing paths leading to victory of opponent.

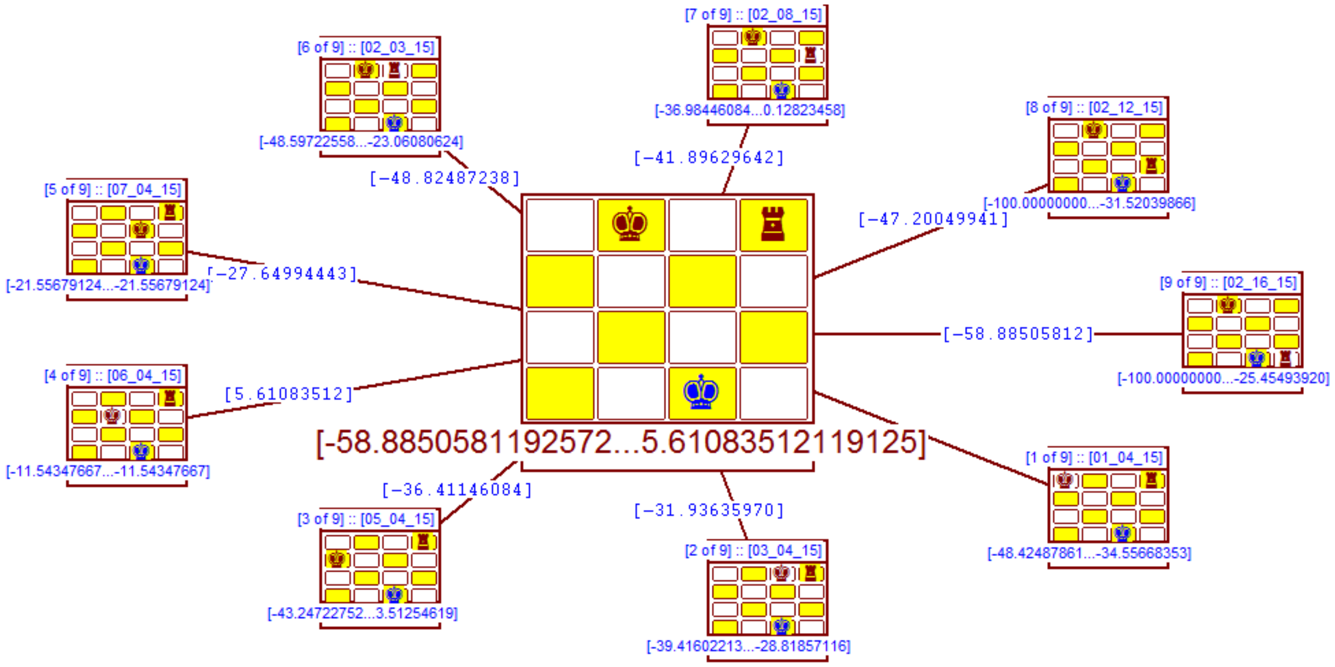


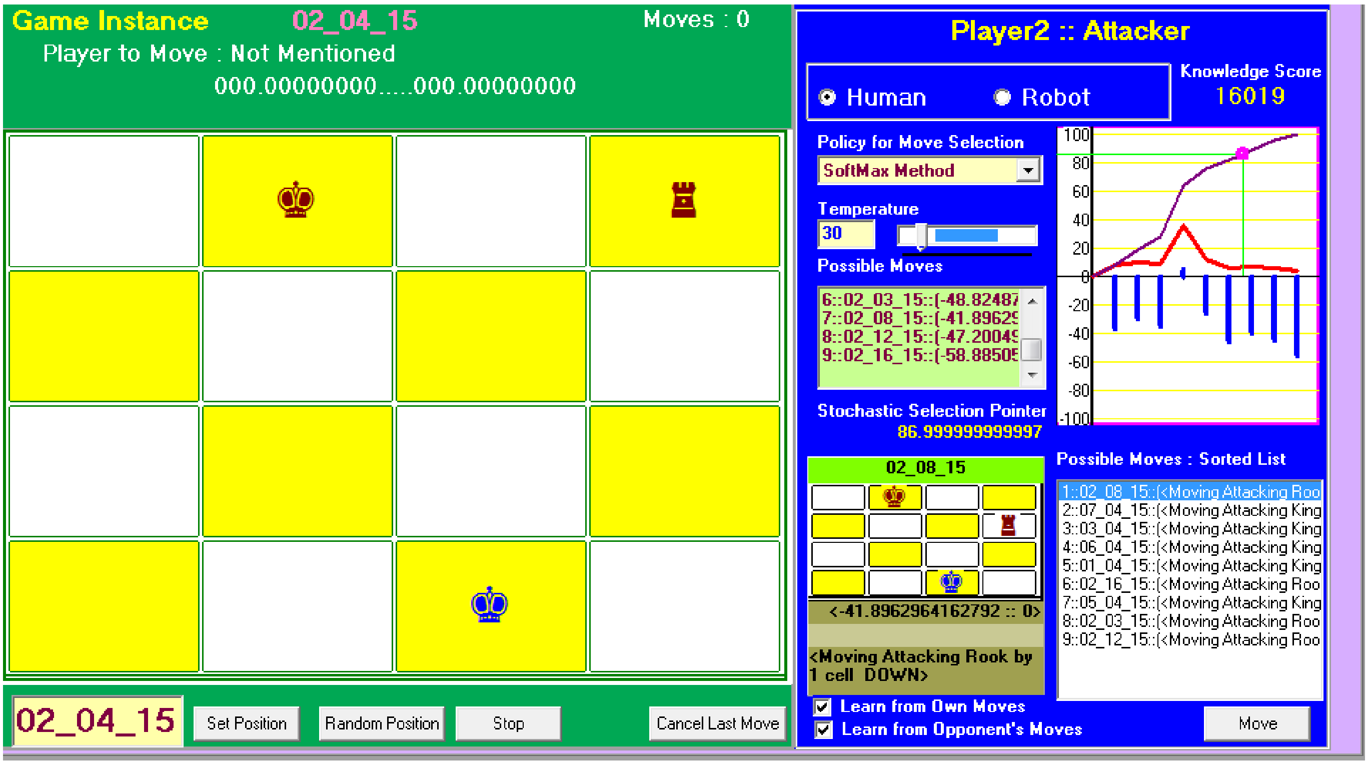
Fig -8: Neighboring states for gaming instance from King_Rook_King End game model of Chess.

The Probability values for each option derived from the Q-values and by considering the environmental parameters are tabulated as follows.

Table -5: Stochastic Selection process based on prior knowledge

Neighboring States		Q-Value	Probabilistic Score	Cumulative Score
1	01_04_15	-38.797003068213	8.28559616377641	8.28559616377641
2	03_04_15	-31.9363597003852	10.4145796886759	18.7001758524523
3	05_04_15	-36.4114608383521	8.97135460412964	27.671530456582
4	06_04_15	5.61083512119125	36.4076849833031	64.0792154398851
5	07_04_15	-27.6499444255134	12.0141752890095	76.0933907288946
6	02_03_15	-48.8248723841513	5.93137640450607	82.0247671334006
7	02_08_15	-41.8962964162792	7.47234466215406	89.4971117955547
8	02_12_15	-47.2004994064437	6.26138909518969	95.7585008907444
9	02_16_15	-58.8850581192572	4.24149910925572	100

A screen shot from the gaming experiment is given below.



Game Instance 02_04_15 Moves : 0
Player to Move : Not Mentioned
000.00000000.....000.00000000

02_04_15 Set Position Random Position Stop Cancel Last Move

Player2 :: Attacker
● Human ○ Robot
Knowledge Score 16019

Policy for Move Selection
SoftMax Method
Temperature
30
Possible Moves
6::02_03_15::(-48.82487
7::02_08_15::(-41.89629
8::02_12_15::(-47.20049
9::02_16_15::(-58.88505

Stochastic Selection Pointer
86.999999999997

02_08_15
<-41.8962964162792 :: 0>
<Moving Attacking Rook by 1 cell DOWN>

Possible Moves : Sorted List
1::02_08_15::(<Moving Attacking Roo
2::07_04_15::(<Moving Attacking King
3::03_04_15::(<Moving Attacking King
4::06_04_15::(<Moving Attacking King
5::01_04_15::(<Moving Attacking King
6::02_16_15::(<Moving Attacking Roo
7::05_04_15::(<Moving Attacking King
8::02_03_15::(<Moving Attacking Roo
9::02_12_15::(<Moving Attacking Roo

☑ Learn from Own Moves
☑ Learn from Opponent's Moves
Move

Fig -9: The Stochastic selection of next move in a gaming instance from King_Rook_King End game model of Chess.

In the above Gaming model, the Stochastic Selection Pointer is a random number in the range 0 to 100. At this particular instance shown in Fig-9 its value is 86.999999999997 which falls in the slot for move 7 (02_08_15). This move is having a negative Q-value which indicates that this is an explorative move by the Software robot which simulates the Attacker in the King_Rook_King end game model. The selection process is architected based on the Pareto principle and can be visualized using a Pareto chart.

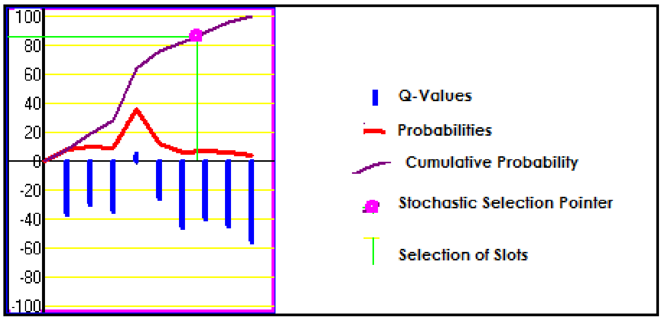


Fig -10: The Pareto Chart for visualizing the stochastic selection process for SoftMax method in King-Rook-King End Game model

7. MODEL BASED APPROACH FOR GENERALIZATION

Most of the Soft Computing Applications involve decision making situations in which the selection of an option from a set of possible options is done based on the knowledge from prior experience and the environmental settings.

The case study mentioned in the Fig-6 makes a selection of the next move based on the following factors.

- Score value of the option based on prior experience
- Policy of the Decision Maker
- Blind (Random Selection)
- Greedy(Maximum Exploitation)
- Soft max(Balanced approach between Exploration and Exploitation)
- Parameters Controlling the Strategy
- Learning Rate (How far the new knowledge is used to update existing knowledge)
- Temperature(For controlling the risk taking levels of the player)

Similarly, according to the specific requirements of the project situation, solutions can be derived using the gaming meta-models.

CONCLUSIONS

Learned Guesses are key characteristics of Intelligent systems. Learned Guesses can be effectively used for implementing many of the decision making models in intelligent systems for real world applications. Statistical tools like Pareto Chart and concepts like Pareto principle can play a significant role in architecting Learned Guesses for Intelligent Software systems based on Soft Computing techniques.

REFERENCES

Own works

[1]. Anil Kumar S and Muralidharan K B, “Rewarding Punishments and Punishing Rewards in Reinforcement Learning” , Proceedings of the National Conference on Recent Trends in Computational Intelligence and Image Processing(RCIP), University of Calicut, pp22-27, February 2016.

[2]. Anil Kumar S, “Wise Optimizations through Reinforcement Learning – Demonstrated with Simple concepts in Gaming and Gamification (WORLDs) ”, Project Thesis submitted at Department of Computer Science, Cochin University of Science and Technology for the fulfillment of MTech Degree Programme in Software Engineering, April 2016.

[3]. Anil Kumar S, “Significance of Gamification and Reinforcement Learning in Indian Language Computing with Creative Works”, Proceedings of the 6th National Conference on Indian Language Computing, ISBN 978-93-80095-75-2, pp 22-27, February 2016.

Other References

[1]. Richard S Sutton and Andrew G Barto, “Reinforcement – An Introduction”,MIT Press, Cambridge, 1998

[2]. Michael Milton, “Head First Data Analysis”, O’REILLY, ISBN 978-0-596-15393-9.

BIOGRAPHIES

Anil Kumar S attained his M.Tech Degree in Software Engineering from Department of Computer Science, Cochin University of Science and Technology, by receiving GATE scholarship and availing leave for higher studies while working in the University of Kerala as Assistant. Earlier, he has completed his Master Degree in Computer Applications from IGNOU and MPhil in Computer Science from Bharathidasan University on part-time basis, along with his career in Software industry in the levels of Programmer and Analyst Programmer. His key areas of research interest include Multi-disciplinary problem solving, Alternative Solution Models and Puzzles. At present he is focusing on research in areas related to Software Engineering and Soft Computing. The author can be contacted at aks.kerala.india@gmail.com