**Predicting movie genre derived from plot description using LDA and LSA modelling techniques**

Guddu Kumar Shah Faculty of Engineering Environment and Computing Coventry University Coventry, United Kingdom shahg4@uni.coventry.ac.uk

*Abstract—This paper seeks to present the study of the two topic modelling algorithms, LSA (Latent Semantic Analysis) and LDA (Latent Dirichlet Allocation). The paper also acknowledges the behavior of these two algorithms on the dataset cons isting of list of movies along with their plot descriptions. The study focuses on applying the topic modelling algorithms on the plot description for predicting the genre of the movie and finding the most popular genre of movies produced from 1901 to 2017 all around the world. The result of both the model are compared on their performances and LDA performed better than LSA on the corpus used.*

Keywords—Topic modelling, Latent Semantic Analysis – LSA, Latent Dirichlet Allocation – LDA.

# **INTRODUCTION**

Topic modelling is a machine learning technique that is capable of scanning large sets of textual documents, identifying pattern in words and phrases and clustering them. It is an unsupervised machine learning technique that counts the number of words and groups them into similar word patterns to generate topics within the documents. The topics are generated by detecting the frequency of most occurring words and calculating the distance between them (Pascual, 2019). Topic modelling is the quickest and easiest way of analyzing data as it doesn’t require training as compared to other supervised algorithms. It is a also inferred as a statistical model for discovering hidden semantics structures in an unstructured documents which intuitively reduces the time taken by human in analyzing the texts and modelling it manually. There are many applications of topic modelling in real world such as opinion summarization, sentiment analysis, Internet of Things, Blockchain, content recommendation, search engine optimization, word sense disambiguation (Fatma, 2019). The motive of this paper is to apply topic modelling on the movie plots description for genre detection. This paper focuses on examining the application of topic modelling on the plot descriptions for movie genre detection. Movies/Films are intuitively the primary sources of entertainment in today’s world. There are thousands of movies produced every year of different genres. The main inspiration behind this paper is to find out which movie genre is the most popularly produced for the audiences. Movie genres are basically a stylistic categorization where a particular movie is defined based on the characters, plots, narration, mood, theme and tone which will influence the techniques of movie/film making and production. There are several movie/film genres like westerns, musicals, thriller, drama, horror, science fiction, comedy, action, film noir, crime and many more. Each of these genres have their distinctive textual representation. Movie plot description plays an integral part in the genre reflection where people can easily evaluate the movie genre information from their plot descriptions. The phrases and sentences inside the plot summaries highly represent the movie genre. People mostly prefer to go through the plot description of movies prior to watching them for getting a brief idea about the movie. This is why, plot summaries is a way for portraying the information of genre to the people. For instance, when the plot summary conveys humorous problems that must be overcome before lovers meet, the movie is probably stating to be a romantic-comedy (Cargal, 2007). With this we can acknowledge that genre information is basically hidden inside the plot description of the movie. This paper seeks to predict the genres of the movies by building two machine learning models considering and modelling genre information represented by every single sentence inside plot description of Wikipedia Movie plots dataset from 1901 to 2017. The dataset contains eight columns The techniques for examining the plot summaries are Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). Both the techniques are evaluated to identify which performs better on detecting the genre. This method will eventually help in movie recommendation by using the Information of plot summaries.

In past, there were several studies done on movie genre classification using audio, visual and textual features from posters, trailers and texts. Rasheed et al. (2005) employed audio and visual features along with motion content, lightening key and average screen shot length to classify movie genres. The study by Zhou et al. (2010) used bag-of-visual-words model along with shot classes for classifying genre. Further, another study by Ekenel et al. (2013) used combination of low level audio and visual features with signal energy, color and textured features for visual representation in genre prediction. This was a great improvement achieved in movie genre classification. Later, Convolutional Neural Network based model was used for genre classification from movie trailer (Wehrmann et. al, 2017).

There were several more studies which employed textual sources included in plots as input to the Support Vector Machine. According to the study by Hong et. al (2015) the textual features were extracted from websites and Probabilistic latent semantic analysis was applied for classifying genre. One more study, represented plot summaries as a Bag-of-Words and used column network on genre classification (Pham et, al. 2017). There was an another innovative approach where Bi- LSTM model was used for genre classification by considering genre represented by single sentence (Ertugrul et. al, 2018). The most recent study used topic modelling on textual plot summaries for investigating key questions in genre, mapping, canonicity and change over time for movie genre prediction (Matthews et. al, 2021).

# **THE DATASET**

The dataset selected for this project was found on Kaggle and is called “Wikipedia Movie Plots”. The dataset contains 34,886 unique records of movies. There are eight features/columns in the dataset for each record. The features are Release Year, Title, Origin/Ethnicity, Director, Cast(main actor and actress), Genre, Wiki Page, Plot(description/summary). The plot descriptions are scrapped form Wikipedia. The movies data collected in the dataset has Release Year from 1901 to 2017. These movies belongs to 21 ethnic groups (American, Australian, Bangladeshi, British, Canadian, Egyptian, Hong Kong, Filipino, Assamese, Bengali, Bollywood, Kannada, Malayalam, Marathi, Punjabi, Tamil, Telugu, Japanese, Malaysian, Maldivian, Russian, South Korean, Turkish). According to the dataset, there are 2265 unique genres of movies recorded. The Wiki page feature contains the corresponding links to the Wikipedia page from where the movies are extracted. The Plot(description/summary) is the categorical value describing the plot information in sentences. The first 5 records from dataset is displayed in the Figure 1 below.

Figure 1: Head of Dataset

Text

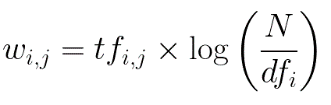
Description automatically generated with medium confidence

# **METHODS**

The methods used in this project as mentioned before are Latent Semantic Analysis and Latent Dirichlet Allocation.

1. LATENT SEMANTIC ANALYSIS

LSA is a topic modelling technique for analyzing the relationship between set of documents and the terms contained within. It is a mathematical method for modelling and simulation of the meaning of words and passages by analyzing corpora representation of natural texts. In other word, it is a method for extraction of relationship between words in text documents. LSA uses a document-term matrix for describing the occurrences of term in documents. It does not require semantic networks, syntactic or morphological parsers dictionaries, grammars as it’s a corpus-based process. The input to LSA model are represented by texts that are divided into smaller parts being a paragraph, a sentence or a document. This method is also stated as tf-idf, where tf stands for term frequency and idf stands for inverse document frequency. The tf calculated the number of times a word appears in a document, this calculation is just relative to the current document not the whole corpus whilst idf is calculated which makes the term relative to the number of documents or the entire corpus. LSA model basically replaces the raw counts in document-term matrix with tf-idf score. The tf-idf assigning weight for term j in document i is shown below.



Here, the input initiates with computed term-doc matrix on a document which is separated into chunks. Singular Vector Decomposition (SVD) are applied to get the most important values by performing dimensional reduction. A latent semantic space is created which only makes the use of most important values. This newly created latent semantic space is further used for computing similarities between words and document. This consider words that are occurring in similar context and relates them semantically.

1. LATENT DIRICHLET ALLOCATION

LDA is a probabilistic machine learning and the most popular topic modelling technique. Here, a document is made up of number of words and each topic has number of words assigned to it. The LDA main objective is to find topics for a document belonging to, on the basis of words in it. LDA generates a cluster assigning keywords to those cluster according to probability : P(word | topics) and P(topics | documents). These probability values are calculated randomly by assigning values at the beginning followed by iterative process by reassigning the probabilities based on topics. The probabilities are recalculated on every iterations until the threshold criteria is achieved. LDA technique is based on the Bayesian framework which allows the model to infer topics on the basis of observed words by making the use of conditional probabilities LDA is a technique that ignores the syntactic meaning of the document and treats the words as bag of words. LDA uses the statistical mixture hypothesis which assumes a corpus having several topics, later, statistical distribution is developed for each topic. Words will move along with the topics on suitability basis of word for topic and also topic suitability for documents. Here, suitability is determined particularly by frequency counts and Dirichlet distribution. Hence, the observed structure of the documents confirms the latent relationship of words and generated latent topic structure. There are two main hyperparameters for LDA, alpha for controlling document similarity and beta for controlling topic similarity. Here, high value of alpha means maximum number of topic being assigned to a document while high value of beta means maximum number of words being used to determine the topic. The basic structure of LDA model is shown in Figure 2 below.

Figure 2: LDA Model

Diagram

Description automatically generated

# **EXPERIMENTAL SETUP**

The experiment as mentioned before is to apply LSA and LDA modelling techniques on the plot description to predict genre and comparing them on their performances. There are number of steps that were followed during this experiment which will be explained in detail. The experiment was performed using Python programming language along with importing of scientific libraries that contained predefined functions. The code is tested on Jupyter Notebook that supports IPython command shell for Python and other programming language. The project runs Python 3.10.5 on Jupyter Notebook 6.4.8 installed on Windows 10 – 64 bit. The entire pachage is installed by downloading Anaconda Distribution framework 4.12.0. The complete code is added in the appendix below. The steps followed are:

1. Loading Libraries
2. Exploratory Data Analysis and Pre-processing
3. Modelling using LSA and LDA analysis
4. **Loading Libraries**

The libraries used for this experiment are pandas, numpy, matplotlib, seaborn, IPython display, sklearn, spacy. These packages helped with loading the algorithm and visualization that were being used for the experiment.

1. **Exploratory Data Analysis and Pre-processing**

The exploratory data analysis starts with loading the dataset in the jupyter notebook. The dataset showed 34886 rows and 8 features. The idea was to represent in a way that can be easily analyzed. On analyzing the frequency of the movie realese year was highest in 2017 as in Figure 3 and American origin was the highest number of movies released and the drama genre was the highest number of movies recorded in the dataset which was 5964.

Figure 3: Release year frequency

Chart, histogram

Description automatically generated

Figure 4: Movies origin distribution

The movie which belonged to more than one genre were dropped and top 6 genres(drama, comedy, horror, action, thriller, romance) were taken and labelled from 1 to 6. These 6 genres are represented according to their frequency in Figure 5.

Figure 5: Top 6 genres

Chart, histogram

Description automatically generatedThe plot summaries/description was analyzed and found that total number of words were 5240143 amongst which 253147 was unique words in the corpus which gives the insight about vocabulary of the dataset. The cleaning of the total number of words in the description which can be achieved by importing stopwords function.The stopwords function were imported from the nltk package for excluding occurrences of conjunctions, prepositions, articles etc. The number of words and unique words drastically dropped as depicted in Figure 6.

Figure 6: Words and unique words counts in plot description

Chart, bar chart

Description automatically generated

A picture containing graphical user interface

Description automatically generated Now in the pre-processing step the feature construction is done where the plot descriptions are represented in some tracable feature space. This is achieved by using CountVectorizer object from sklearn library yielding n\*k document-term matrix where k in number of distinct words and n is descriptions/summaries in the plot. The frequency distribution of top 10 words in plot summary for the 6 genres are presented in Figure 7.

Figure 7: Top 10 words in plot summary in 6 genres

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

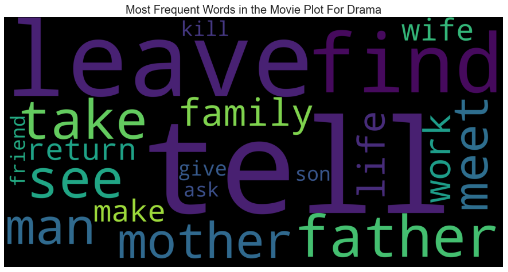
Description automatically generated

Chart, bar chart

Description automatically generated

The worldcloud is imported to visualize top 20 most frequent words in all 6 genres Figure 7.

Figure 8: Most frequent words in 6 genres

A picture containing text

Description automatically generatedText

Description automatically generatedText

Description automatically generatedText

Description automatically generatedText

Description automatically generated

Now all the plot descriptions are taken into one corpus and upon further pre-processing the unique numbers of words obtained were 37585 and then the dictionary of words were created by from genism library. The bag-of-words were generated from the dictionary, here I obtained very high rank and sparse data which is now ready to implement clustering algorithm. The algorithms to be used are LSA and LDA, both of them takes document-term matrix as input and produce n\*N topic matrix as output, where N is number of topics.

1. **Modelling using LSA and LDA analysis**

* Latent Dirichlet Allocation

With respect of LDA,after creating the dictionary for term frequency calculation, the lda model was built using genism function and selecting 6 as the number of topics and having symmetric alpha, this means that the number will neither be high nor low. The figure below Figure 8 shows the generated topics with their corresponding words.

Figure 9: Topic generated by LDA model

A picture containing text

Description automatically generated

After the experiment, we found that LDA topic modelling could divide the plot into well defined topics. Each topic generated resembled to the corresponding genre as predicted. This is clear that LDA performed well in separation of the generated topics into their particular genre. The coherence score was 0. The topic coherence score for this model was 0.5203 after successful modelling. The figure Figure 9 shown below displays the topic modelling visualization for the each of the 6 genres from the plot corpus.

Figure 10: Topic modelling visualization for LDA

Chart, bubble chart

Description automatically generated

* Latent Semantic Analysis

Now, after performing the experiment of topic modelling using LSA model, the topic matrix gave the predicted topic of each plot description. The same pre-processing steps were applied for this modelling. The topics generated by this technique is shown below in Figure 9.

Figure 11: Topics generated by LSA model

A picture containing text

Description automatically generated

The LSA divided the topics corresponding to their genres as shown after applying this technique. The genre classification according to the topic generated could be easily done. The coherence score obtained by this model was 0.3736.

# **RESULT**

During analysis of the dataset it was found that the drama genre of movies were produced or releases the maximum from 1901 to 2017 which was 5964. During the exploratory data analysis several stopwords were introduced along with the word tokenization that helped in making the words easier that would help the model easy for analyzing. The cleaned text generated resulted in lemmatized text. The result was the best version for modelling on LDA and LSA topic modelling techniques for genre prediction. Both the techniques performed well on modelling the plot summaries from the dataset. The dictionary created contained 37585 unique tokens. The LSA technique was simply implemented over the corpus and dictionary along with 6 topics(0: drama, 1: comedy, 2: horror, 3: action, 4: thriller, 5: romance). This After applying list of topics were generated which clearly represented the corresponding genres. Same mechanism was used for LDA topic modelling. Coherence score is the measure represented by the highest probability of the words that particularly belongs to the topic (Zvornicanin, 2021). Coherence score for LSA was 0.376 and for LDA was 0.5203 also illustrated in Table 1. This clearly represents that LDA works better in creating and assigning the words to the topic. Hence, on this basis we can evaluate that LDA modelling techniques performed better than LSA in performance aspect.

Table 2: Coherence Score

|  |  |
| --- | --- |
| Modelling Technique | Coherence Score |
| Latent Semantic Analysis | 0.376 |
| Latent Dirichlet Allocation | 0.5203 |

# **SOCIAL, ETHICAL, LEGAL ISSUES**

The issues faced during working on this project was that the genre features of some of the movies were unknown, which was difficult to evaluate so it was dropped for further analysis. Also, there were genre of movies that were the combination of more than two specific genres. The movie genres with more multiple genres combined was very difficult to evaluate and to a single genre. With this, finding the most popular genre of movies released over the past 116 years from 1901 to 2017 was not easy. The dataset seemed to biased to one particular origin of movie according to the dataset which is far different from reality. The problems occurred during analysis was the processing time of some of the libraries and functions. There were more than 2000+ classes(genre) to be predicted in this project. Handling such huge amount of classed would have been very difficult for the machine learning techniques. Machine learning methods are not capable enough to handle those volume of classes. It was a bit difficult to measure the accuracy of the model as there is no concept of train test split concept in topic modelling. However, topic modelling is the best method for analyzing huge corpus and find patterns or topics. The final issue faced was the processing time for cleaning the plot description when stopwords was used and collecting separating every words.

# **DISCUSSION AND CONCLUSION**

Several works were done in past for movie genre prediction using topic modelling which achieved great results. Genre prediction from plot description helps a lot in building the movie recommendation system for the audiences. Movies plot summary gives basic information about the casts, location, ethinicity and genre. These details give audiences the idea about the movie in totality. People need to put time in obtaining these information. With a movie recommendation system people can get all the information about the movies. Genre prediction from the textual representation seems to be a bit more hectic than from trailers and posters. It is because lots of things are needed to consider to build the model. There may be presence of ambiguous characters, unknown languages which are needed to be interpreted in a way making it into a machine readable form. LDA topic modelling is the best method for genre prediction as from the experiment and after number of iterations the model generates more great results. It is undeniable that LDA always perform better in large datasets compared to LSA and PCA. It is because it performs class separability along with dimensionality reduction for documents.

To conclude, in this paper, I performed movie genre prediction from plot summaries using LDA and LSA technique. Instead, of making the entire plot summary as input, it was divided into sentences and then into words for developing a bag-of words in order to perform topic modelling tasks. Results show that LDA outperforms LSA. It can also be seen that the performance of LDA is better than the LSA from the experiment.

##### **REFERENCES**

Cargal, T. B. (2007). *Hearing a film, seeing a sermon: Preaching and popular movies*. Westminster John Knox Press.

Ekenel, H. K., & Semela, T. (2013). Multimodal genre classification of TV programs and YouTube videos. *Multimedia tools and applications*, *63*(2), 547-567.

Ertugrul, A. M., & Karagoz, P. (2018, January). Movie genre classification from plot summaries using bidirectional LSTM. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)* (pp. 248-251). IEEE.

Fatma, F. (2019, April 4). *Industrial applications of topic model.* Medium. <https://medium.com/@fatmafatma/industrial-applications-of-topic-model-100e48a15ce4>

Hong, H. Z., & Hwang, J. I. G. (2015, June). Multimodal PLSA for movie genre classification. In *International Workshop on Multiple Classifier Systems* (pp. 159-167). Springer, Cham.

Matthews, P., & Glitre, K. (2021). Genre analysis of movies using a topic model of plot summaries. *Journal of the Association for Information Science and Technology*, *72*(12), 1511-1527.

Pascual, F. (2019, September 19). *Topic Modelling: An Introduction*. MoonkeyLearn. <https://monkeylearn.com/blog/introduction-to-topic-modeling/>

Pham, T., Tran, T., Phung, D., & Venkatesh, S. (2017, February). Column networks for collective classification. In *Thirty-first AAAI conference on artificial intelligence*.

Rasheed, Z., Sheikh, Y., & Shah, M. (2005). On the use of computable features for film classification. *IEEE Transactions on Circuits and Systems for Video Technology*, *15*(1), 52-64.

Wehrmann, J., & Barros, R. C. (2017, April). Convolutions through time for multi-label movie genre classification. In *Proceedings of the Symposium on Applied Computing* (pp. 114-119).

Zhou, H., Hermans, T., Karandikar, A. V., & Rehg, J. M. (2010, October). Movie genre classification via scene categorization. In *Proceedings of the 18th ACM international conference on Multimedia* (pp. 747-750).

Zvornicanin, E. (2021). *When Coherence Score is Good or Bad in Topic Modeling?.* Baeldung*.*

##### **APPENDIX**

* Link to the github repository containing the complete code.

<https://github.com/guddu095/7135CEM_Task1.git>

**Fuzzy Logic Optimized Controller for an Intelligent Assistive Care Environment**

*Abstract*: The purpose of this paper is to implement Fuzzy logic controller for a smart flat by giving inputs to control the outputs by developing mathematical rules using Mamdani Inference system. The system was built using MATLAB Fuzzy Toolkit. After the successful development of the Fuzzy Logic Controller, the optimization was performed using Genetic Algorithm in Global Optim Toolbox and it was compared to other optimization algorithm for accuracy. These evolutionary algorithms were very useful for improving performance of Fuzzy Logic Controller of a room.

*Keywords: A smart flat, Mamdani Fuzzy Inference System, Genetic Algorithm*

1. **Fuzzy Logic Controller Design**

The design of Fuzzy Logic Controller was developed using MATLAB Fuzzy Logic Toolbox. The Fuzzy logic takes input, transfers it to the fuzzy logic inference and produces output. The design of smart flat using FLC is shown in Figure 1.

Figure 1: FLC Design for smart flat

Diagram

Description automatically generated

The workflow of the FLC starts with taking the number of inputs and setting up the parameters according to the interest. Then, the values of the parameters were set for Input Membership Function and Output Membership Function. Further, the set of rules are constructed for using in the Fuzzy Inference. After that we used the model called Mamdani fuzzy model. Later on, once the number of inputs were feeded, fuzzification was done on the input . The data was sent to the fuzzy inference. Finally, the defuzzification was done using ‘Centroid ’ technique to obtain the output.

1. **Input and Membership Functions**

There were five inputs which were taken to design the FLC for a smart flat. These inputs are Temperature, Humidity, Light, Time of day and Activity. At this point, the class were provided. For each input the membership functions are defined. The membership are either trapezoidal or triangular here.

1. **Temperature**

Temperature was divided into 5 measure sclaes. These were VL for Very Low, L for Low, M for Moderate, H for High and VH for Very High. The trapezoidal membership function was taken for VL and VH whilst for L, M and H triangular function were taken.

|  |  |  |  |
| --- | --- | --- | --- |
| Temperature Membership Function | | | |
| Sr No. | Membership  Function | Range | Type |
| 1 | VL | -10°C to 10°C | Trapezoidal |
| 2 | L | 0°C to 20°C | Triangular |
| 3 | M | 15°C to 27°C | Triangular |
| 4 | H | 20°C to 38°C | Triangular |
| 5 | VH | 35°C to 50°C | Trapezoidal |

Diagram

Description automatically generated with medium confidence

1. **Humidity**

Humidity of a flat should range from 0% to 100%. Variation in humidity were measured on the scale from VeryLow(VL), Low(L), Moderate(M), High(H) and Very High(VH). Similar to the temperature VL and VH are trapezoidal and L, M and H are triangular membership function.

|  |  |  |  |
| --- | --- | --- | --- |
| Humidity Membership Function | | | |
| Sr No. | Membership  Function | Range | Type |
| 1 | VL | 0.1% to 20% | Trapezoidal |
| 2 | L | 25% to 40% | Triangular |
| 3 | M | 45% to 60% | Triangular |
| 4 | H | 65% to 85% | Triangular |
| 5 | VH | 75% to 100% | Trapezoidal |

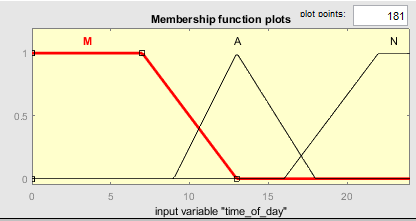
Diagram

Description automatically generated

1. **Time of day**

This input is divided into three membership functions.These are Morning (M), Afternoon(A) and Night(N). Here M and N are taken as trapezoidal and A as triangular.

|  |  |  |  |
| --- | --- | --- | --- |
| Time of day Membership Function | | | |
| Sr No. | Membership  Function | Time (Hours) | Type |
| 1 | M | 00:00-13:00 | Trapezoidal |
| 2 | A | 09:00-18:00 | Triangular |
| 3 | N | 16:00-24:00 | Trapezoidal |



1. **Activity**

The external movement or activity done by user is ranked as Low (L), Medium (M) and High (H).Here, L and H membership functions are taken as trapezoidal while M as triangular.

|  |  |  |  |
| --- | --- | --- | --- |
| Activity Membership Function | | | |
| Sr No. | Membership  Function | Time (Steps/30 minutes) | Type |
| 1 | L | 0-15 | Trapezoidal |
| 2 | M | 10-40 | Triangular |
| 3 | H | 30-55 | Trapezoidal |

Chart, line chart

Description automatically generated

1. **Light**

The light levels are divided into 5 different categories. They are Very Low (VL), Low (L), Medium (M), High (H) and Very High (VH). The level VL and VH are trapezoidal whilst L, M and H are taken as triangular. The range is measured in Lux.

|  |  |  |  |
| --- | --- | --- | --- |
| Light Membership Function | | | |
| Sr No. | Membership  Function | Range | Type |
| 1 | VL | 0.00 Lux–4000 Lux | Trapezoidal |
| 2 | L | 3000 Lux–8000 Lux | Triangular |
| 3 | M | 7000 Lux–12000 Lux | Triangular |
| 4 | H | 10000 Lux–16000 Lux | Triangular |
| 5 | VH | 16000 Lux-20000 Lux | Trapezoidal |

A picture containing diagram

Description automatically generated

1. **Output and Membership Functions**

The environmental parameters to be controlled are cooling fan, heater, blinds and dimmer switches.

1. **Cooling fan**

The speed of cooling fan is divided into 5 different categories. These are Very Low (VL), Low (L), Medium (M), High (H) and Very High (VH). The VL and VH are trapezoidal whereas L, M and H are triangular.

|  |  |  |  |
| --- | --- | --- | --- |
| Cooling fan Membership Function | | | |
| Sr No. | Membership  Function | Range | Type |
| 1 | VL | 0.0 μm to 5 μm | Trapezoidal |
| 2 | L | 3.0 μm to 12.5 μm | Triangular |
| 3 | M | 10 μm to 15 μm | Triangular |
| 4 | H | 14 μm to 20 μm | Triangular |
| 5 | VH | 18 μm to 25 μm | Trapezoidal |

A picture containing line chart

Description automatically generated

1. **Heater**

The heating inside the flat is categorized into 5 membership functions, Very Low (VL), Low (L), Medium (M), High (H), Very High(VH). VL, VH are trapezoidal and L, M, H are triangular.

|  |  |  |  |
| --- | --- | --- | --- |
| Heater Membership Function | | | |
| Sr No. | Membership  Function | Range | Type |
| 1 | VL | 0 to 2 | Trapezoidal |
| 2 | L | 1.8 to 4 | Triangular |
| 3 | M | 3 to 6 | Triangular |
| 4 | H | 5 to 8 | Triangular |
| 5 | VH | 7 to 8 | Trapezoidal |

A picture containing line chart

Description automatically generated

1. **Blinds**

Blinds are categorized as Half Closed (HC), Closed (C), and Open (O). The trapezoidal function is used for membership in O and C. The triangular function is used for HC.

|  |  |  |  |
| --- | --- | --- | --- |
| Blinds Membership Function | | | |
| Sr No. | Membership  Function | Range | Type |
| 1 | O | 0.5 to 1 | Trapezoidal |
| 2 | HC | 0 to 1 | Triangular |
| 3 | C | 0 to 0.5 | Trapezoidal |

**Chart, line chart

Description automatically generated**

1. **Dimmer Switches**

For dimmer switches, five levels of switching were done, VL (Very Low), L (Low), M (Medium), H (High) and VH (Very High). The membership function of L, M and H are triangular while trapezoid for VL and VH.

|  |  |  |  |
| --- | --- | --- | --- |
| Dimmer Switches Membership Function | | | |
| Sr No. | Membership  Function | Range | Type |
| 1 | VL | 0.00 Lux–4000 Lux | Trapezoidal |
| 2 | L | 3000 Lux–8000 Lux | Triangular |
| 3 | M | 7000 Lux–12000 Lux | Triangular |
| 4 | H | 10000 Lux–16000 Lux | Triangular |
| 5 | VH | 15000 Lux-20000 Lux | Trapezoidal |

Line chart

Description automatically generated with low confidence

1. **Fuzzy Inference System**

Fuzzy inference System plays the major role in designing Fuzzy Logic Controller. It is the CPU of the fuzzy system. Here, FLS uses fuzzy set theory for mapping input to output. For constructing FLC, Mamdani Inference was used as the inferential framework. Here, several set of input variables are feeded in to generate one corresponding output.The GUI of Fuzzy system is shown in Figure 3. This system determines each Input Classifier and initial cross value of the variables. After that this method uses the membership functions in the input set for determining smallest intersection point among all the computed data. Then this membership values are used as a input into the classifier of output varaibles. The graphical representation of each rule’s effects on the corresponding output variable is obtained. With each itiration, the membership value of the obtained result is the average of the membrs values determined for all rules.

Figure 2: Fuzzy Inference System

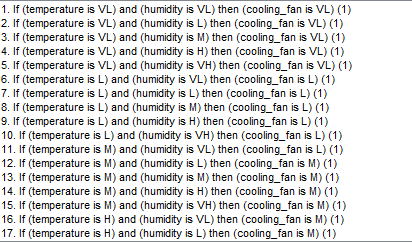
Diagram

Description automatically generated

1. **Fuzzy Rules**

The rules generated inside the rule ediotor of the Fuzzy Logic System is shown in the figure Figure 2 below.

Figure 3: Rules

A close-up of a document

Description automatically generated with low confidenceA picture containing text, newspaper

Description automatically generatedA picture containing text, newspaper

Description automatically generatedText

Description automatically generated

1. **Defuzzification**

This method is used to get the final crisp output from the input set and fuzzy rules. The defuzzification is achieved by using ‘Centroid’ approach, which is the summation of all the centre points and area of the result. This is done by the Mamdani Induction in the numerator and adding area to the denominator.

1. **Justifying FLC design**

The temperature plays an intrinsic part in adjusting the operations of cooling fan and heater. For dimmer switch, the fine-tuning is done with the help of input light intensity. External properties like user’s activity and time of day are integrated into the decision making of blinds, whether it should be open, half closed or closed. The inputs are basically the output detectors throughout the flat. The values of temperature, humidity and input light are splitted into 5 fuzzy sets for more specific regulation. The time periods are divided into three time shifts. With this choices we assume that the settings fits the objectives very well. High temperature value dampness requirement have role in fuzzy rules selection. Therefore, cooling fan has high value and heater has low value and vice-versa. The high light intensity light in the morning triggers the dimmer swith to set lower than night’s low intensity light. When the user’s activity. According to this framework, we chose Mamdani Deduction and center of area defuzzifier that are mostly used. Furthermore, the min-max approach utilized in Mamdani Inference for and/or inference aggregate helps in building an inexpensive and user-friendly FLC design.

1. **Output Behaviour**

The output behaviour for the Fuzzy Inference computation with following input set are represented by Figure 4. The inputs taken were 20 degree celsius, 50 percent humidity, user activity of 10 cadences at 4 o’clock in the evening and 5000 lux of light.. The results obtained are as given below. Cooling fan 12.5, heater 4.5, blinds 0.5 and dimmer switch illuminance as 94600 Lux.

A picture containing table

Description automatically generatedA picture containing text, window blind

Description automatically generatedGraphical user interface, application

Description automatically generated

Plots of input versus output can be shown in figure below.

Figure 5: Temperature, Humidity vs Cooling Fan Figure 6: Temperature, Humidity vs Heater

Chart, surface chart

Description automatically generatedChart

Description automatically generated

Figure 7: Time\_of\_day , Activity vs Blinds Figure 8: Time\_of\_day, Light vs Dimmer\_switch

Chart, surface chart

Description automatically generatedChart, surface chart

Description automatically generated

1. **Genetic Algorithm for Optimization**

1. Fitness Function

Function y =fitnessfreak(x)

m(1) = -70\*x(1)+100\*x(2);

m(2) = 3.5\*x(1)-2.5\*x(2);

m(3) = 116-559\*(x(3)/1000)-60\*(x(4))^2;

m(4) = 110\*x(4)+x(6);

y = 10000/(m(1)+m(2)+0.5\*m(3)+m(4));

end

Parameter values for 5 input variables are x(1), x(2), x(3), x(4) and x(5). The relative importance of each input is used for fine tuning the weights assigned to parameters m(1), m(2), m(3) and m(4).

1. Population

Here, we have chosen a Double Vector population. The starting population is as follows: [-20 -20 0 15 5000 ;-10 0 0 6 0;0 20 17 5 4000;-5 10 13 10 2000;10 30 0 15 5000;20 50 0 10 8000;10 35 0 0 5500;20 60 17 .5 10000;35 85 0 0 13500;20 55 0 0 10000;30 75 0 0 14000;40 90 20 10 18000;35 85 17 .5 17000;50 100 21 40 20000;60 100 20 50 25000]. Each element in the vector represents a parameter for membership function on input. The total number of population is set to 200.

1. The Rank is used as a scaling function and selection function is chosen to be Roulette. Mutation function is Adaptive Feasible. Elite Count is calculated as 0.05\*Population size.Finally the crossover fraction which is selected from the best population obtained is 0.833. The figure below represent the result obtained.

Graphical user interface, application

Description automatically generated

1. **CEC’2005 functions for comparison between optimization techniques**

The Genetic Algorithm and Particle Swarn Optimization are the two approaches which will be used for exploration. Shifted Schwelf’s problem 1.3 with bounds [-100 100], f bias = -500.Shifted Rastrigin’s function with bounds [-5,5] and parameters f bias = 320. The following results were obtained.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Particle Swarm Optimization | | | | | | |
|  | CEC Function = 2 | | | | CEC Function = 9 | |
|  | **D = 2** | **D = 10** | | | **D = 2** | **D = 10** |
| **No.** | **Local Minima** |  | **Local Minima** | **Local Minima** | | **Local Minima** | |
| 1 | -500.45 |  | -498.9817 | -329.005 | | -274.2745 | |
| 2 | -500.564 |  | -499.5687 | -329.654 | | -300.1489 | |
| 3 | -500.645 |  | -499.554 | -329.546 | | -267.3478 | |
| 4 | -500.456 |  | -499.654 | -329.645 | | -305.5456 | |
| 5 | -500.456 |  | -500.5454 | -329.654 | | -322.0435 | |
| 6 | -500.5458 |  | -549.9945 | -329.54 | | -320.0353 | |
| 7 | -500.4564 |  | -500.5456 | -329.4564 | | -327.0526 | |
| 8 | -500.545564 |  | -500.4564 | -329.546 | | -311.33256 | |
| 9 | -500.65465 |  | -499.99748 | -329.54 | | -327.03526 | |
| 10 | -500.4654 |  | -499.998 | -329.45 | | -327.326 | |
| 11 | -500.541 |  | -500.45 | -329.41565 | | -325.032145 | |
| 12 | -500.546 |  | -499.9459 | -329.456 | | -326.02231 | |
| 13 | -500.546 |  | -499.95899 | -329.456 | | -318.035215 | |
| 14 | -500.54164 |  | -499.95649 | -329.456 | | -310.13225 | |
| 15 | -500.465465 |  | -499.9659 | -329.4564 | | -327.0545 | |
| Max\_Value | -500.5464 |  | -499.9327 | -329.5665 | | -310.1545 | |
| Min\_Value | -500.65454 |  | -550.456 | -329.4156 | | -328.0256 | |
| Mean\_Value | -490.546 |  | -499.95589 | -329.545 | | -322.4415 | |
| Standard Deviation | 10.654e-06 |  | 9.2461e-05 | 0.353456 | | 5.7210 | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Simulated Annealing** | | | | | | |
|  | **CEC Function = 2** | | | | **CEC Function = 9** | |
|  | **D = 2** | **D = 10** | | | **D = 2** | **D = 10** |
| **No.** | **Local Minima** |  | **Local Minima** | **Local Minima** | | **Local Minima** | |
| 1 | -500.45 |  | -498.9817 | -329.005 | | -274.2745 | |
| 2 | -500.54 |  | -499.5687 | -329.654 | | -300.1489 | |
| 3 | -499.9459 |  | -499.554 | -329.546 | | -267.3478 | |
| 4 | -500.565 |  | -499.654 | -329.645 | | -305.5456 | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 5 | -490.45 |  | -499.5415 | -329.545 | -277.546 |
| 6 | -490.58 |  | -497.82515 | -330.465 | -286.894 |
| 7 | -499.9789 |  | -499.3542 | -328.987 | -280.546 |
| 8 | -499.9459 |  | -498.5456 | -329.889 | -301.1546 |
| 9 | -500.456 |  | -499.4454 | -329.999 | -296.1548 |
| 10 | -499.9454 |  | -498.8545 | -329.999 | -273.2654 |
| 11 | -499.9450 |  | -498.25436 | -329.999 | -305.1545 |
| 12 | -499.543 |  | -498.68456 | -329.999 | -296.1545 |
| 13 | -499  .98 |  | -499.6954 | -329.999 | -317.546 |
| 14 | -499.878 |  | -499.1157 | -329.999 | -292.1947 |
| 15 | -499.564 |  | -499.5545 | -329.999 | -270.2954 |
| Max\_Value | -499.954 |  | -497.8564 | -329.999 | -267.345 |
| Min\_Value | -499.8787 |  | -499.66547 | -329.999 | -317.0465 |
| Mean\_Value | -499.9659 |  | -499.04565 | -329.999 | -289.4654 |
| Standard\_Deviation | 9.3242e-05 |  | 0.5346 | 0.6367 | 15.104556 |

It can be observed that Particle swarn optimization outperforms the genetic algorithm techniques. The code for the same is attached in Figure 9 below.

Figure 9: Code

Text

Description automatically generated

Text

Description automatically generated