ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 13, No. 1, 2018, *pp*.003-021

Efficient Approach for Land Record Classification and Information Retrieval in Data Warehouse

C.B. David Joel Kishore [1] and T. Bhaskara Reddy [2*]
[1] *Research Scholar, Department of Computer Science, Rayalaseema University, Kurnool, Andhra Pradesh, India.*
[2] *Professor, Department of Computer Science & Technology, Sri Krishnadevaraya University, Anantapur, Andhra Pradesh, India.*
(Received October 21, 2017, accepted December 18, 2017)

Abstract. A data warehouse collects the recent and ancient data that are used for creating analytical reports and put together to produce useful information. Retrieve the information accurately from a large source of data is a challenging task. A novel ANN-FUZZY-CSO approach is proposed to predict and retrieve the information accurately. First, the artificial neural network (ANN) classifies the input data for ordering the information to construct a database as different classes. Then, the mongo database will store a large amount of data for facilitating easy maintenance, prompt updating of land records and security. After that, the optimized fuzzy ranking function is used to retrieve the information from the database based on the optimal fuzzy rules using cat swarm optimization algorithm. The fuzzy rules provide a ranking for an individual field in the database. The accurate results for the user query are retrieved using cat swarm optimization (CSO) algorithm. The optimized fuzzy rules allow the users for easy access to their records. Finally, the performance is evaluated for the retrieval results

Keywords: artificial neural network, cat swarm optimization, data warehouse, fuzzy, mongo database

1. Introduction

A data warehouse is a structure utilized for the data analysis and storing the amount of information. Also, the data warehouse is a central repository which can be placed all data relevant to the management of the organization and which provides the prominent information to effectively manage the organization [1]. The information stored in the warehouse is in the scheme of derived views of data from the sources [2]. The data warehouse defines how data's are composed, merged, elucidated, trained, and that should be training-based managerial information [3]. They can determine the facts that facilitate managers to analyze situations requiring action and to know the situation and its intention. They must be established that authorize a manager to place and utilize the relevant organizational knowledge which predicts and scope the impact of a declaration added. These are the significant challenges for the research community in data warehousing [4]. For finer declaration, a data warehouse is processed to keep the large database.

The big data technologies used for storing an infinite number of data and we make a good decision [5]. So, they proposed remarkable uses of the bid data technology in the data warehouse. Big data is a major technology to analyze the large volume of data and extract the necessary information or knowledge for future action [6]. As a rule, the knowledge extraction process must be extremely effective and near real-time on the grounds that putting away all watched information is about hopeless. Also, the big data application using the large data sets, has a complex structure which is trouble of storing, analyzing and visualizing the results are varied [7-10]. The action of research into big amounts of data to reveal invisible patterns and private correlations named as big data analytics. These data's are essential to improve achieving richer and deeper insights and getting an advantage over the competition in the companies or Institutions. For this reason, big data implementations are a demand to be analyzed and completed as exactly as possible.

In existing, the large-scale real-world taxi trace data of a big city with a population of 6 million that are utilized to store in big data technologies. There the big data improved clustering algorithm (iterative DBSCAN) is presented to extract regions, respectively to the quality of the data. Six kinds of features are designed, four classifiers are combined into the evaluation. The achievement of different features and classifiers is evaluated, and the best feature combination is also achieved. But the system considers only current addresses regions with pure land use that does not consider regions with multiple land-use classes. So the Information of the passenger cannot be accessed clearly [11]. Another one sales forecasting for utilizing fuzzy logic and Naïve Bayesian for large data. There, the sales forecasting for any product is

'Poor', then it gives hint that some preventive measures like discount or offer should be taken to enhance the sale. If sales forecasting for any product is 'Good' then it gives hint that enough amount of stock should be made available in the store. So the system having low preparing time to given poor accurate results [12]. A fast Messy Genetic Algorithm used in an Advanced Hybrid Genetic Algorithm with support vector machine (SVM) system was presented. The early prediction of conflict disposition in the initial stage of public-private partnership projects using fmGA. The system providing a term of accuracy, precision, sensitivity, specificity, and area under the curve. But the system has some classification problem so that doesn't give the efficient result [13].

Another current technique is the land cover classification with better resolution remote sensing information coordinating worldly elements extricated from time arrangement coarser resolution information. The coarser resolution vegetation index information is initially intertwined with better resolution information to acquire time arrangement better resolution information. The coarser resolution vegetation index data is first combined with finer declaration data to obtain the time series. The fleeting elements are removed from the intertwined information and added to enhance classification accuracy. But the classification accuracy is lower than the other classification techniques [13]. Another one the big data using service-oriented decision support systems. The system given better results such as reduction in unit service costs due to increase in operational size (scale), reduction in unit service costs due to increase in number of services being developed and provided (scope) and reduction in unit costs as a result of increase in number of benefits are put through supply/demand chain (speed). But the result is not accurate and does not efficient [14].

The major contribution of the proposed work:

1) A classification model is proposed for gathering the facts to arrange a database concerning land profit, grazing method, and land use.
2) A novel optimized fuzzy based ranking approach for facilitating easy maintenance and prompt updating of land records.
3) The Cat Swarm Optimization (CSO) algorithm will allow the users to easily access their records optimally.
4) The mongo database used in our work will make the land records secure.

The remaining paper is organized by the following sections. Related work is described in section 2. Section 3 defines the proposed methodology in detail. The Experimental result is discussed and comparison of various methods are explained in section 4. Finally, Section 5 provides the conclusion.

2. Related Work

Torsten Priebe *et al.* [15] have proposed a combined methodology to structure and describe business requirements in large data accelerated projects, e.g. data warehouse employment, that precise and explicit data definition suitable to further accession and assignment of data governance responsibilities. The information can be placed in the center of the business model. The Data Administrator gave the information with the help of end-to-end analysis, design, development, testing to data quality checks. In addition, they display that the method is proper beyond conventional data warehouse platform. Also, the big data landscapes and information science initiatives are applied, the wherever business necessities analysis is commonly ignored.

Vikas S Shah *et al.* [16] have proposed an approach in consideration of multi-faced compliance-aware data services (DS) that enables the degree of distinction in business rules. They presented an approach to continuously monitor regulatory updates and rationalization to translate them into CS. The research also presents runtime structure to evolve and govern compliance-aware DS. The categorization and corresponding implementation of CS into the DS are identified and implied in association with business rules. Formulae to evaluate the degree of distinction and assessment criteria to monitor governance of deployed compliance-aware Data Services are illustrated with an example implementation and validated in the number of actual deployment iterations.

Osden Jokonya *et al.* [17] developed and validated an IT adoption structure to comfort organizations with IT adoption administration. The framework useful for improving IT adoption administration in organizations as it addresses different concerns during IT adoption such as stakeholder buy-in, discrimination, oppression, and stakeholder participation. The structure promoter for the need to understand the problem context before selecting suitable approaches for intervention. On that note, the properties of a

structure to promote IT adoption administration in organizations has to be comprehensive in nature to address the complexity of IT adoption decision-making in organizations. Therefore, there is no one-size-fits-all method to IT adoption administration in organizations.

Veerendra Kumar Rai *et al.* [18] have proposed a three-layered typological classification of the methods and guidelines commonly occupied in a common project. They have granted a topological classification of the policies for IT project management based on their related positions on a five-dimensional framework. This is one of the various attainable methods for classification. Also, they proposed an exploratory approach where they have attempted to classification and its usability in better governance through an Agent-Based Model. The work can be taken further to improve upon the classification and have it validated through empirical analysis. It provides more robust study by presenting a taxonomy based classification of the IT governance policies.

Lukasz Golab *et al.* [19] have proposed some data warehouses, which fuse the benefits of old data warehouses and data stream systems. In existing model, external sources push insert-only data streams into the warehouse with a huge range of inter-arrival times. Then they proposed a scheduling structure that handles the complexity preserved by a stream warehouse that contains view hierarchies, priorities, data flexibility, failure to anticipate updates, diversity of update jobs caused by different inter-arrival times and data strength between different sources, and transitory overload. Finally, they provide a collection of update scheduling algorithms and comprehensive simulation experiments to map out factors which affect their performance.

Mehdi Kashfi *et al.* [20] presented that five different kinds of data warehouse architectures such as independent data mart, centralized data warehouse, dependent data mart, homogeneously distributed data warehouse and heterogeneously distributed data warehouse and afterward a comparison plan will be described. First, the role of data warehouse and afterward, all sorts of centralized and distributed architectures were described. Consequently, the research proposed distributed data warehouse architecture with high compatibility with optimal architecture for organizations so that the organizations' data processing system will be compatible with its ever-growing data and fast.

Jinglan Zhang *et al.* [21] described that Big Audio Data which analyzed and managed for Environmental Monitoring. The system Environmental monitoring was enhancing difficult as human activity and weather change place greater pressures on biodiversity, leading to an improving the needs of data to generate good decisions. Acoustic sensors could gather data crosswise over vast ranges for amplified periods making them alluring in environmental checking. Be that as it may, overseeing and breaking down extensive volumes of environmental acoustic data was an extraordinary test and was subsequently upsetting the viable use of the big dataset gathered. It describes an overview of our recent techniques for gathering, storing and analyzing large volumes of acoustic data efficiently, precisely and cost-effectively.

Bingwei Liu *et al.* [22] proposed that Big Data Analysis used with Naive Bayes Classifier. The system used this technique was easy and the whole system for sentiment mining on huge datasets using a Naive Bayes classifier with the Hadoop framework. The NB classifier can measure the data easily, even without a database. A run of the mill strategy to get important information was to extricate the sentiment or opinion from a message. Machine learning technologies were broadly utilized as a part of sentiment characterization as a result of their capacity to "learn" from the training data set to anticipate or support decision-making with generally high precision. Although dataset was extensive, a few algorithms won't scale up well so the system calculates the scalability of Naive Bayes classifier in wide datasets. The NBC was able to determine the opinion of million reviews increasing the throughput.

Lisette *et al*. [23] have proposed a multidimensional data model that integrates bias data separated from client analysis setting into the collective data warehouse. The process of extracting sentiment data considers as assessments products made by customers on both products features and produces a rich data set which enables the complex queries. As an outcome, another vital aftereffect of this work comprises of demonstrating the helpfulness of MDX queries over the incorporated data warehouse to complete the requirement of the voice bounding market.

Jemal H *et al.* [24] have presented that Security of Big Data used with Large Iterative Multitier Ensemble Classifiers which particularly customized for big data. These classifiers were large, however very simple to produce and utilize. They could be large to the point that it seemed well and good to utilize them just for big data. They were produced naturally as an after effect of a few emphases in applying ensemble Meta classifiers. They join various ensemble Meta classifiers into a few levels at the same time and consolidated them into one consequently created an iterative framework. So that the numerous Ensemble

Meta-Classifiers work as an indispensable section of another Ensemble Meta-Classifiers at higher levels. The structure portrayed the exhaustive examination of the implementation of LIME classifier for a problem concerning the security of big data.

Ahmad Taher and Aboul Ella [25] have made up the Linguistic Hedges Neuro-Fuzzy Classifier with Selected Features (LHNFCSF) was shown for decreased dimensionality, highlight choice, and order that implementation of Four real-world datasets using with Neuro-Fuzzy Classifier. The new classifier was contrasted and alternate classifiers for various arrangement issues. The outcomes demonstrated that applying LHNFCSF lessens the measurements of the issue, as well as enhanced arrangement execution by disposing of excess, clamor tainted, or insignificant features. However, the existing technique not just decreased the dimensionality of expensive data that likewise can accelerate the calculation time of a learning calculation and disentangle the grouping errands. To overcome the difficulties in existing works we provide a new framework with trending technology. By using the proposed technology accuracy of the system can be improved. The following section elaborates the proposed work.

3. Proposed ANN-FUZZY-CSO for land data classification and information retrieval

The land records data of the Andhra Pradesh Government can be gathered that act as the input of the proposed system. Then, the collected land record data can be classified by using Artificial Neural Network Classifier. The Artificial Neural Network classifies the input record data for an individual person of the AP Government. The Artificial Neural Network performs the classification using the different process to give the efficient results. Then, the classified information saved in the mongo database. Mongo database is the document-oriented database so the individual information is stored in an effective manner. After that, the information is efficiently retrieved by using the optimized fuzzy ranking function. It is used to retrieve the user-specified information. The below Fig. 1 displays the planning of the projected system.

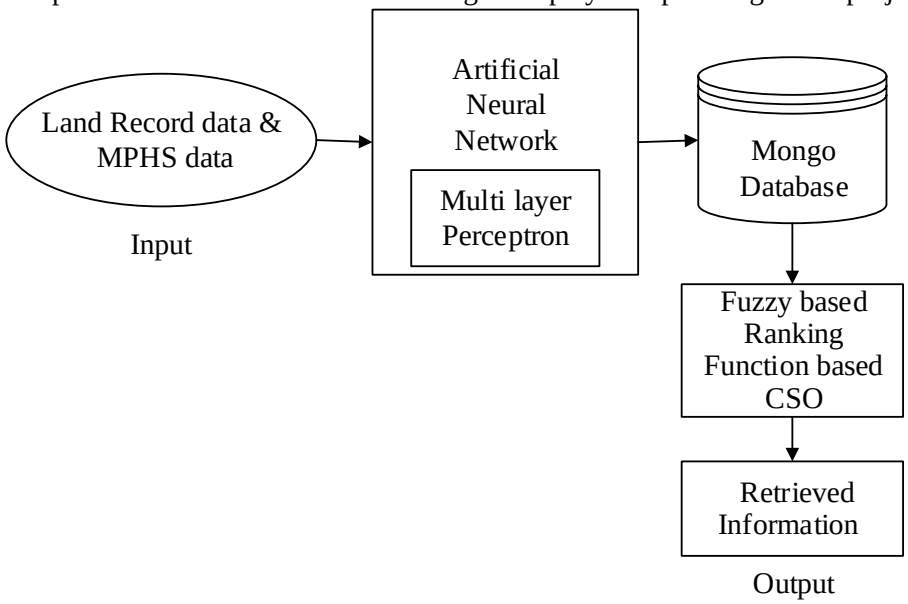

Fig. 1: Architecture of our proposed system.

The main modules of the proposed system are Data collection of land records, Classification of the information, database storage, and Information retrieval.

3.1. Data collection for Land Record Information

The Data collection is the main work for the proposed system. To achieve better data collection process Multi-Purpose Household Survey (MPHS) data and the land records data is collected from different sources. Land Records is collected depend on the following process. State level land records are collected by the land record departments and Settlement Commissioner. Also, the information of particular District land records is gathered by the District Inspector as well as Taluka Inspector. A function of the Land Records department is described in the following paragraphs.

1. An Up-to-date survey, classification and settlement records are collected and protected by creating field operations. It allows updating the changes in survey records.
2. By utilizing Analytical information, all data dependent on the land records are collected and maintained effectively.
3. For preserving survey and target land records are the dispute in civil courts. It simplifies the methods and reduces the cost.
4. Record of Rights can be supervised by periodical monitoring, maintenance, and reconstruction of the boundary marks of unique fields.
5. Compensation operations are performed to achieve periodical enhancement.
6. Appropriate preservation and pervasive scale are organized to obtain village site surveys.
7. Up-to-date changes are performed in village survey to manage the records.
8. All tahsil maps are reprinted and separated under various departments. It provides easy access records to the user. Finally, the concatenation of the revenue officers in survey and settlement matters are taking place.

Our proposed system utilizes Multi-Purpose Household Survey (MPHS) data and the land records data gathered from the different sources. The information present in the record is converted into the uniform format. Then the collected information used as input for the classifier to classifies the information.

3.2. ANN Land Record Classification

In our proposed system, we are using Artificial Neural Network (ANN) to classify the information. During the data collection, distinct types of data are to be collected. But the information is not in proper order. So the user required information cannot be accessed easily. ANN is a supervised machine learning algorithm that has the benefit for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, each data item will be specified as a coordinate in n-dimensional space where n is the number of aspects of the information. The Artificial Neural Network is the best classification process that separates the land record information as per the user convenience.

ANN is an effective calculating system, which acquires the huge accumulation of parts and that units are interconnected between some function. These units are referred to as node which operates in a simple process like parallel function. The main objective of ANN is to promote a system to act various computational tasks such as pattern recognition and classification, optimization and data clustering.

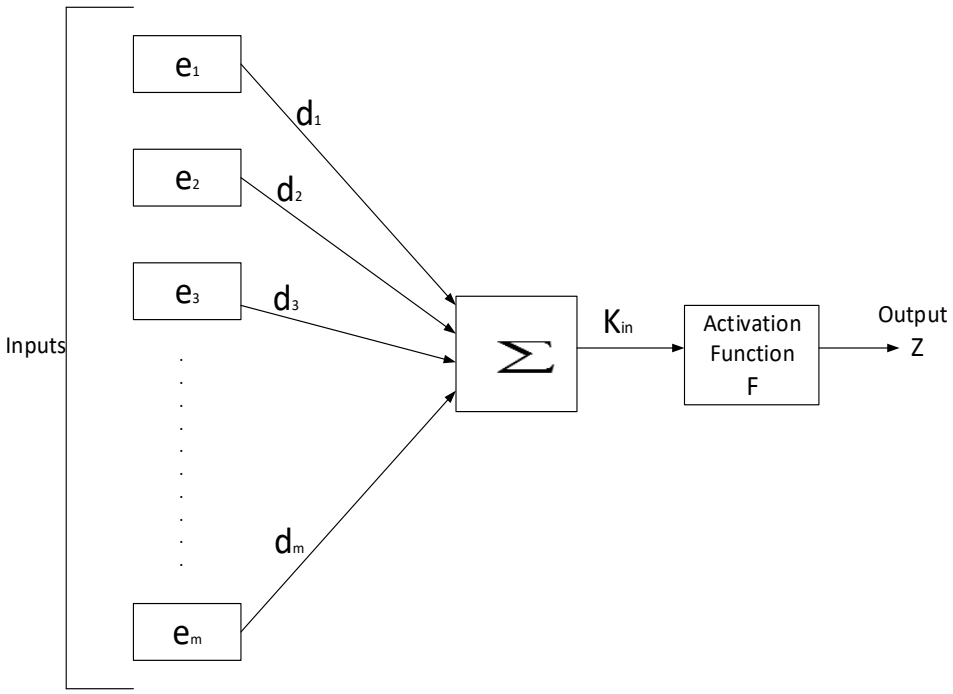

Fig. 2: Model of Artificial Neural Network.

Fig 2 shows the model architecture of ANN. First, the collection of input data $e_1, e_2 \ldots\ldots e_m$ is multiplied by connection weights $d_1, d_2 \ldots\ldots d_m$. Then, the products are summed and fed through the activation function.

The output of Summation is given,

$$K_{in} = \sum_i^n e_i . d_i \tag{1}$$

The output of the activation function as,

$$Z = F(K_{in}) \tag{2}$$

This implementation is possible with other network structures as the different transfer function. The multilayer perceptron is also called as deep feed forward neural networks which map the set of input data to appropriate set of outputs. It exists of the diverse region of nodes in a supervised graph with each region related to the succeeding one. Fig. 3 shows the process of the multilayer perceptron. It has three types of layers such as input layer, hidden layers, and the output layer.

1) Input layer: number of nodes depends upon the number of inputs which receive the data from land records.
2) Hidden layer: these regions come from the input layer and the output layer which receives the data from the input region. The function is to encode the input and maps it to the output.
3) Output layer: Input flows from the input layer and passing hidden layers to the output layer. Error correction of neural weights is approved in the contradictory area. This is approved by the back propagation algorithm. Mainly, the classification based on the size of the output layers.

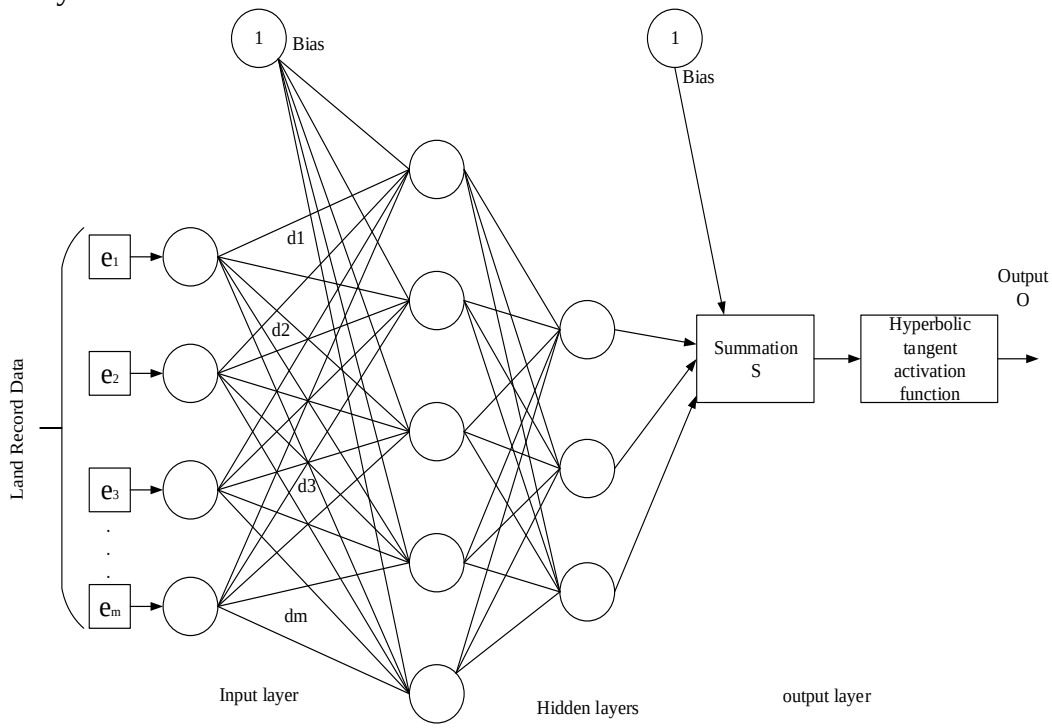

Fig. 3: Process of Multilayer Perceptron.

Summation process $S = \sum_i^n e_i . d_i$ bias can be considered as,

$$S = \sum_{i=1}^n e_i . d_i + BIAS . d_i \ , \qquad \text{when BIAS=1}$$

The activation process can be used in different types such as the sigmoid activation function and hyperbolic tangent activation function which means multilayer perceptron is a sigmoidal activation in the form of a hyperbolic tangent. It is a real mathematical work that denotes for all actual input values and has the non-negative derivative at

each point. The hyperbolic tangent is a ratio of corresponding hyperbolic sine and hyperbolic cosine function via. Hyperbolic tangent activation function,

$$0 = \tanh(s) = \frac{\tanh(s)+1}{2} \tag{3}$$

The back propagation algorithm can be used for weight changes in hidden layers and output layers. This type of algorithm uses the delta-rule which computes the deltas such as local gradients of each data going backward direction to reaches the input layers. An Error can occur in network situation that is the difference between desired output and network output. The delta rule is needed in weights to match the desired output. Adjust the weight for delta rule which adding the current weights to produce a new weight for the output layer. The difference between current weights and previous weights are multiplied by momentum coefficient.

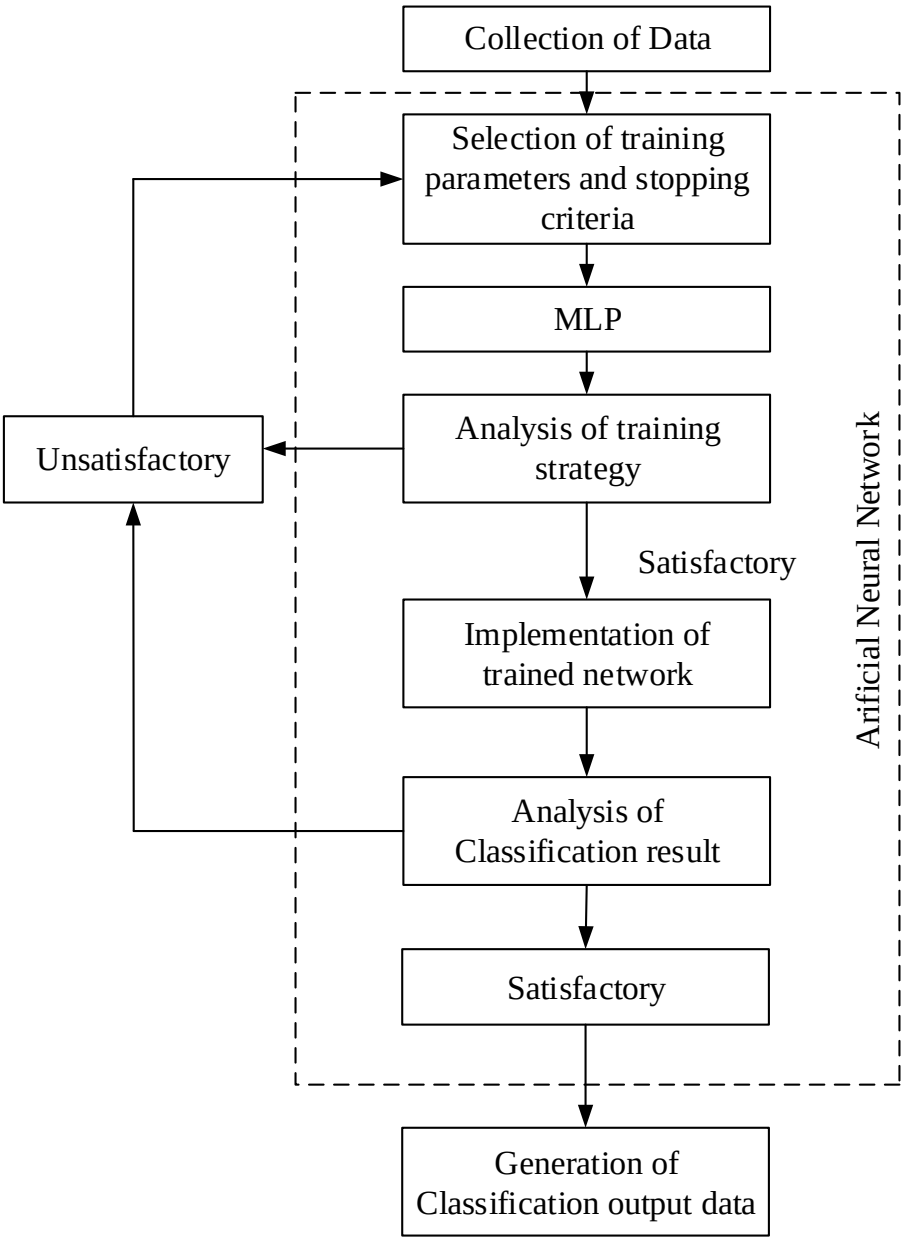

Fig. 4: Flow Chart for the Artificial Neural Network Classifier.

The classification of Artificial Neural Network based on the subset of feed-forward ANN which can be used to data excavating scheme. This network is used to identify the particular patterns and classify them into specific groups. The feed forward ANN is simple and output is decided by its reviewing of input and the power of the network. In supervised learning, the mapping function is used to input variables to output

variables because when you have a new input data which predict the output variables for that data. It makes the predictions of data and corrected by the algorithms so achieves acceptable performance. Fig. 4 shows the flow chart for the classification of ANN. Data can be selected for particular parameters and train the data in the specific network. The trained data can be analyzed and implemented data if satisfactory, to give the classify output, whether unsatisfactory the process can be done in again and again. The Artificial neural network, which has the input layer, output layer, and two or more trainable weight layers (consisting of Perceptron) is called multilayer perceptron or MLP.

Table 1: Process of input classification

No	Region	Name	mail-id	property	district
1	Kadapa	Parthasarathy	sarathy888@gmail.com s^2 0 feet	0 acre0yard	1
2	Kadapa	Raju Reddy	raju222@gmail.com s^2 0 feet	0 acre0yard	1
3	Chenguta	Vijayalekshmi	viji1234dev@gmail.com s^2 0 feet	0 acre0yard	1
4	Narasapur	Venkadanathan	venkat8400@gmail.com s^2 0 feet	0 acre0yard	2
5	Guntur	Srihari Krishna	sriharikrish@gmail.com s^2 0 feet	0 acre0yard	2

In Classification section, data can be collected from Multi-Purpose household system which acts as the input of the classification. Table 1 shows the input classification table includes the different district of the Andra Pradesh Government, landholders and their properties information. Each of the districts can be collapsed itself. So ANN classifier for classifies the data and each of the districts will be separated. Each of the individual information is separated using Multilayer Perceptron in ANN classifier for user convenience.

Table 2: Process of output classification

No	Region	Name	mail-id	property	district
1	Kadapa	Parthasarathy	sarathy888@gmail.com	0 acre0yard s^2 0 feet	2
2	Kadapa	Raju Reddy	raju222@gmail.com	0 acre0yard s^2 0 feet	1
3	Chenguta	Vijayalekshmi	viji1234dev@gmail.com	0 acre0yard s^2 0 feet	1
4	Narasapur	Venkadanathan	venkat8400@gmail.com	0 acre0yard s^2 0 feet	2
5	Guntur	Srihari Krishna	sriharikrish@gmail.com	0 acre0yard s^2 0 feet	1

After performed the classification process, each field of the data will be separated in Table 2. The classified data is nothing but the number of users with the region, and their email ids, property, and district. Then, the classified data can be sent to the mongo database.

DOI https://doi.org/2024-JICS-22459 | Generated on 2025-04-15 10:49:18

3.3. Mango Database

In this framework using mongo database for stores the classified data from ANN classifier. Also, the mongo database is the document dependent database. So the separable data can be stored in each field of the database for user convenient retrieval. Mongo database is a database that stores huge objects or files. There is no issue to stores the large quantity of data in the database. The database has a lot of advantages for storing the files in different fields. So the user can access the information from the desired field of the database.

Mongo DB Data Model: Mongo Database is a record database. Each database includes a group of records. Each record can be dissimilar with a variable number of fields. The size and content of each document can be different from each other. The mongo database having rows and columns that don't require to have a schema determined beforehand. Also, the Mongo DB environments are very scalable. The benefit of the Mongo Database is a shot-platform, document-oriented record that gives high-level performance and availability also it provides better scalability. Mongo DB works on the method that based on the collection of the document. The storage architecture of mongo database is shown in Fig. 5.

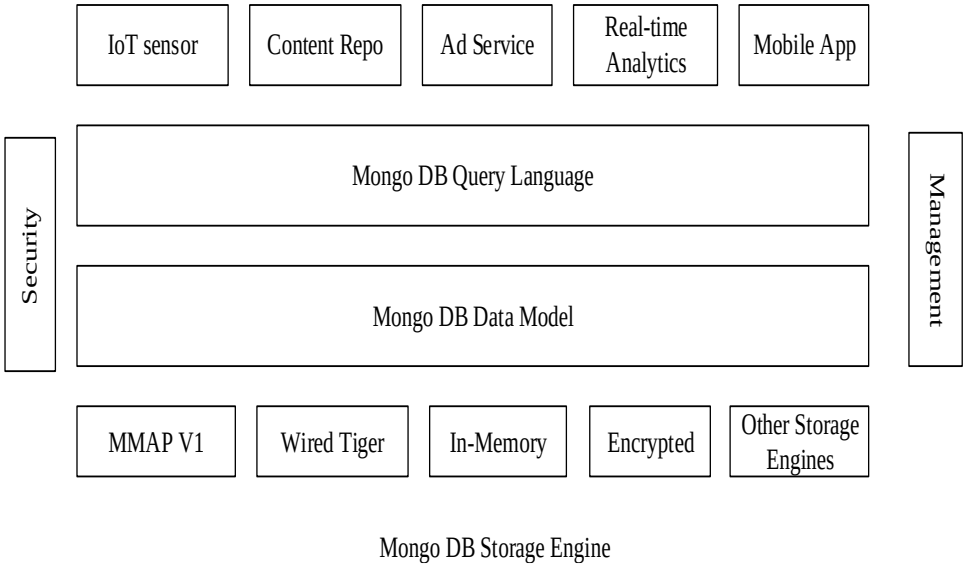

Fig. 5: Storage Architecture of Mongo Database.

Data as Documents: Mongo DB records contribute to the number of the document in a single database because in a relational database information is normal extension across many tables. In a relational database, the data structure would compose multiple rows and columns in a table such as a user, articles, comments, tags, and sections. In Mongo DB the information could be modeled as two collections that are both users and articles. In each blog document that contains multiple comments, multiple tags, and multiple sections, each section expressed as the fixed array. In this work the land record information stored in the database based on the particular format. That includes the username, desired land information, location of the land, length of the land, amount of the land stored in each field of the database. Data as documents is uncomplicated for developers, faster for users. In our structure, the classified data can be stored in mongo database as rows and columns also Individual information stored in each field in the database. So the information can be retrieved from mongo database in particular format. But the user needs information will be retrieved from the database using fuzzy based ranking function techniques that briefly described as given subsection.

Mapping Coordinator: The mongo database includes mapping coordinator for storing and retrieving the data in an effective way. The Mapping Coordinator is also responsible for maintaining and implementing standards for hardware and software to assure compatible, reliable and user-friendly

operation. Also, it handles upload/download, updates of information for user convenience. The major usage of Mapping Coordinator retrieves the information from the database in a secure manner. If the user wants any authenticated information about any organization or user the mapping coordinator to check whether it is authenticated or not. After confirmation, it will provide the information for the required user. It provides more security to the user.

3.4. Information Retrieval for fuzzy ranking based on CSO

The relevant and irrelevant information can be retrieved with the help of fuzzy ranking function using Cat Swarm Optimization process.

Fuzzy ranking function: Fuzzy Based Ranking Function is presented in our proposed system. The fuzzy ranking function provides an acceptable way to retrieve the information using fuzzy logic rules. It operates using a series of if-then statements, which format rules are given below,

IF premise (antecedent) THEN conclusion (consequent)
IF premise1 (antecedent) THEN conclusion (consequent)

The number of data can be placed in fuzzy rules which required relevant and irrelevant information ordered by using weight values w. The membership function of fuzzy is based on the relevant document, irrelevant document, and highly relevant document. The membership value between 0 and 1. The moderate threshold value can be represented by G. Here, input variable is weight value and the output variable is retrieved information.

If $w > G$, then relevant document,

If $w < G$ then irrelevant document,

But, the high relevant document is obtained by using optimization process. Ranking functions recovered a ranking value for all rows in a separation. Depend on the process that is utilized, some rows might collect the similar value as another row. A Ranking is a process of ordering the results of the query according to the measures relevance to the corresponding user. The ranking process will occur in the relevant documents ' c ' from the retrieved results of query ' p '. The ranking process can be obtained using a weighting function $(tf \times idf)$, which conveys how relevant document ' c ' is for query ' p '. Let us assume the set of documents 'C' in mongo database with relevance numerical weights values $b_i = \{b_1, b_2 .... b_n\}$ where $b_i \in \{0,1\}$ as normalization interval which prompts a relevance documents c_i. Here '1' is the maximum relevance weight value corresponding to 'high relevant' and '0' represents 'irrelevant value'. The matrix representation for the Mongo document can be represented as,

$$c = \begin{matrix} c_{11} & c_{12} & c_{13} & \cdots & c_{1n} \\ c_{21} & c_{22} & c_{23} & \cdots & c_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ c_{M1} & c_{M2} & c_{M3} & \cdots & c_{Mn} \end{matrix}$$

For rank retrieved documents according to the value of the weights, then the documents with relevance value b_i will show the ranked list. The rank list for the retrieved documents are given according to the weight value b_i can be calculated as

$$w = (tf \times idf) \tag{4}$$

Where, tf is the term frequency in the query document pair and the fraction of retrieved documents idf can be represented as

$$idf = \log(ni/M)$$

Where, ni is the number of documents retrieved and M is the total number of documents.

The Ranking function E consists of an ordered set of ranks. Each rank consists of relevance weight value $b \in \{0,1\}$ where b represents the relevance numerical weights of the retrieved documents. Each retrieved document is assigned with an ascending rank number.

$$E = \{(1, b_1), (2, b_2), (3, b_3) ..... (n, b_n)\} \text{ where } b_1 > b_2 ..... b_n$$

DOI https://doi.org/2024-JICS-22459 | Generated on 2025-04-15 10:49:18

CSO optimization for High relevant data: The cat swarm optimization based on swarm intelligence optimization algorithm which imitates the action of cats like move fast when they are seeking some intention, and curious about all kind of moving things. It composed in two models such as tracking mode and seeking mode which represents two different procedures in the algorithm. It has finer attainment that finds the global best solutions than another existing optimization algorithm. For tracing mode, models have the behavior of cats when running after a target. In seeking mode, models have the behavior of cats when resetting and observing their environment. The CSO algorithm has the finer attainment in function to minimize the problems are compared to other same optimization algorithms. The CSO has the following section such as, M-dimensional position, velocity for each dimension, fitness value and seeking/tracing flag.

CSO Algorithm: Figure (6) shows the flowchart of Cat Swarm Optimization. The following steps are required in the CSO Algorithm.

Step 1: Create N-cats in the process. The N-cat represents the relevant documents from the result of the fuzzy-based ranking system. Let us assume, query vector P as $P=[p_1, p_2....p_n]$ and the document vector can be represented as C.

Step 2: Digitize the positions of each Cat in M-dimensional space and randomly gives the values in the order of highest velocity. Digitize the position matrix for N-cats $(N\times M)$, where M is the number of variables to be optimized and values are in the range (0, 1). Then initialize the velocity matrix $(N\times M)$ with the range of (0, 1)**.**

Step 3: It represents the criteria of our goal, only need to remember the best position of the best cat (x_{best}). Evaluate the fitness value of each N cats and the best fitness value of cat acts as the global best (g_{best}). Mixture Ratio (MR) dictates the joining of seeking mode with the tracing mode. Mixture Ratio is nothing but most of the time cat is spent in seeking mode.

Evaluate the documents according to their fitness function $f(x)$, keep the best fitness value. Calculating the relevance weight is an important procedure for retrieving the document but here our main goal is to retrieve the high relevant document. To provide a ranked list to the user according to their information requests. Relevance weight is according to the documents. Calculate the weighted root mean squares (RMS) by using average mean weight to determine the fitness value of retrieved documents with respect to the query.

$$f(x)=\max\left\{\frac{1}{l}\sqrt{\sum_{j=1}^{l} w_j^2}\right\} \qquad j=1,2,.....n$$

Where, l- retrieved fuzzy results

Step 4: **Seeking mode operation**: Move the cat according to their modes, if cat_k is in seeking mode, apply the cat into the seeking mode process, otherwise, it applies the tracing mode process.

Seeking mode has the following factors such as the Seeking memory pool (SMP), Counts of dimension to change (CDC), Seeking range of selected dimension (SRD). The seeking memory pool has the number of copies of a cat to be produced in seeking mode. The Count of dimension to change has the out of M-dimensions of a cat if CDC=1, then the dimensions of CDC is randomly changed. The SRD has changed the magnitude of that dimension which is the maximum difference between the new and old value in the dimension selected for mutation. Repeat the procedure for all cats, and evaluate the fitness value of each position. The best-fitted cat is retained and the remaining will be discarded. Repeat it for all seeking mode cats.

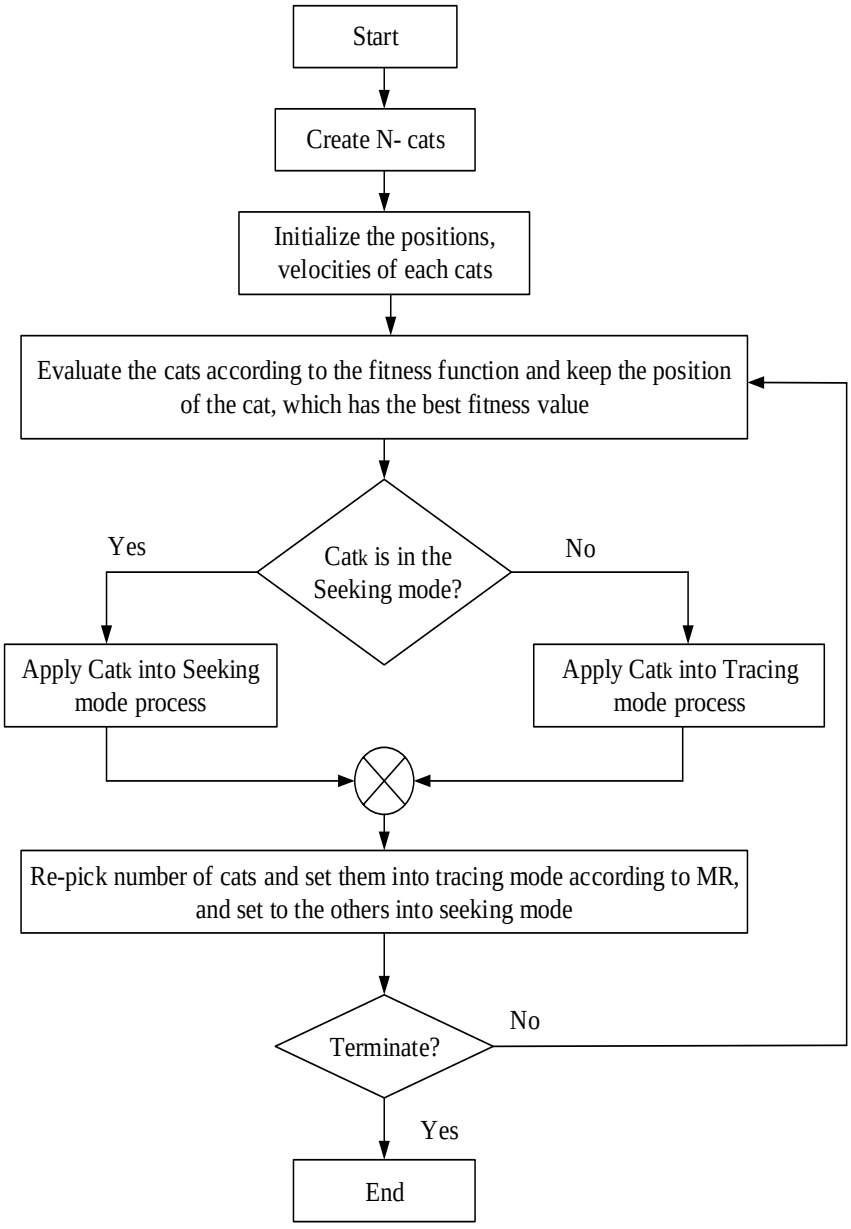

Fig. 6: Flow chart of Cat Swarm Optimization.

Step 5: **Tracing mode operation:** It follows Particle swarm optimization steps without using the personal best (p_{best}) values. Estimate the fitness profit of each cat and find the global best position of these cats. Update the velocity matrix which means update the velocity of each cat, using initial positions and g_{best} values.

$$V_{k,d} = V_{k,d} + r_1 \times c_1 \times (x_{gbest,d} - x_{k,d}) \tag{5}$$

Where, $x_{best,d}$ – Position of the cat, who has the best fitness value, and $V_{k,d}$ is the velocity of cat_k on the d^{th} dimension, $x_{k,d}$ is the Position of the cat, which random variable r_1 in the range of (0, 1) and constant c_1.

$$x_{k,d} = x_{k,d} + V_{k,d} \tag{6}$$

Step 6: Re-pick the number of cats and set them into the tracing mode according to MR, then set the other cats into the seeking mode.

Step 7: Check the termination condition, if satisfied terminate the program, otherwise repeat the steps 3 to step 7.

DOI https://doi.org/2024-JICS-22459 | Generated on 2025-04-15 10:49:18

The ranking functions are nondeterministic. The ranking function is the method of ordering the fuzzy numbers for the data. In this work, the ranking function provides the best rank for an individual field in the database. So the user provides request data can be accessed depends on the ranking function. For example: if any user request the state q information means the ranking function searches which field of the information required to be accessed and provides the relevant information from the database depends on the rank of these fields.

4. Experimental Results and Analysis

In this division, we demonstrate the evaluation of our proposed system. In our experiments aim to classify the different types data and the user can be retrieved the information about land records in Andhra Pradesh government in anywhere. We proposed the classification technique provide the better performance to classify the land records of Andhra Pradesh government. Then the Classification results are stored in the efficient mongo database. After that, the user required information is retrieved from the mongo database using the Fuzzy ranking function. The experimental results and analysis are performed in MATLAB tool. Evaluates the following terms such as precision and recall, F-measure, efficiency, accuracy, and scalability.

4.1. Precision and Recall

Precision is the fraction of the retrieved documents that are relevant and Recall is the fraction of the relevant documents that are retrieved. Precision and recall are denoted by percentage. Precision and recall in information retrieval, suppose we have 100 documents in mongo database, the user can give the query "username". In 100 documents, suppose 20 documents are related or relevant to the query term "username" and other 80 documents are not related or relevant to the username. Now suppose we search the "user age" returns the 25 documents, where 15 documents are relevant and remaining 10 documents are irrelevant. The diagram of precision and recall in information retrieval as shown in fig 9.

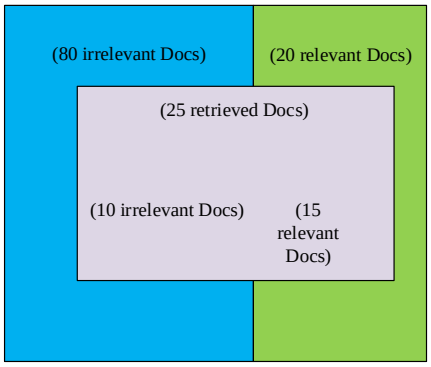

Fig. 7: Diagram for precision and recall in information retrieval.

Precision and recall can be calculated as,
Recall = retrieved relevant / total relevant,

$$p = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

Information retrieval system has Recall (R) and Precision (P) on a collection of test document as per the need of an information. The below Table 3 is expressed in the type of documents.

Table 3. Matrix form of documents.

	Relevant Documents	Irrelevant Documents
Documents Retrieved	True Positive	False Positive
Documents Not Retrieved	False Negative	True Negative

4.2. F-measure

F-measure is defined as the weighted harmonic mean of its Precision and Recall.

$$F \frac{1}{\alpha \frac{1}{F} + (1-\alpha)\frac{1}{R}} \tag{9}$$

Where, $\alpha \in [0,1]$– weight, F-measure is high, only when the both Recall and Prediction will be high. F-measure assumes the values of the interval (0, 1). If F-measure is 0, no relevant documents have been retrieved. If F-measure is 1, all relevant documents have been retrieved.

The Balanced F-measure is denoted by F_1 which have equal weights of Precision and Recall. When $\alpha = \frac{1}{2}$

$$F_1 = \frac{2(\mathrm{Pr}ecision.\mathrm{Re}call)}{\mathrm{Pr}ecision + \mathrm{Re}call} \tag{10}$$

4.3. Accuracy

A group of data points from an order of estimation, the group data is accurate if the values are terminated to the normal value of the number will be determined. While the group data is precise if the values are terminate to the true value of the number being determined. The high accuracy requires both high precision and trueness. Accuracy is the proportion of true results (relevant retrieved and irrelevant not retrieved) among the total number of documents in the database. Based on the contingency matrix, it is expressed mathematically as in below,

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{11}$$

The Accuracy is the summation of Recall and True Negative Rate. The True Negative Rate is simply called as Specificity which denoted by TNR.

$$TNR = \frac{TN}{TN + FP} \tag{12}$$

The existing systems are required the following functions such as Precision, Recall, F-measure, and Accuracy.

Table 4: Performance Evaluation of the Proposed System.

Authors	Method	Precision (%)	Recall (%)	Accuracy (%)
Shulayeva *et al.* [26]	Naïve Bayesian Multinomial Classifier	87	89	85
Chou *et al.* [13]	GASVM Classifier	94.67	91	89.30
Liu *et al.* [22]	Naïve Bayes Classifier	84	87	81
Proposed	ANN-Fuzzy CSO	**97.44**	**98.28**	**98.85**

Table 4 shows the performance evaluation of the proposed work in terms of precision, recall, and accuracy with the existing approaches proposed by several authors.

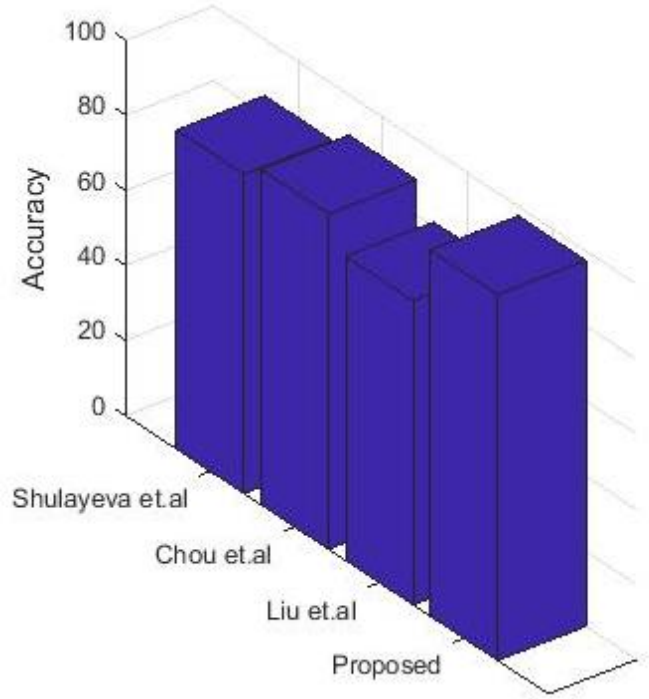

Fig. 8: Performance of Accuracy for Land record data.

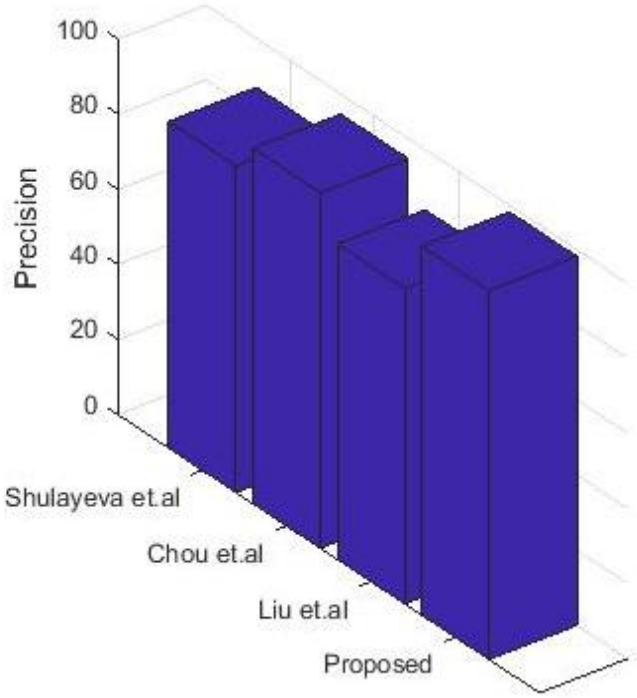

Fig. 9: Performance of Precision for Land record data.

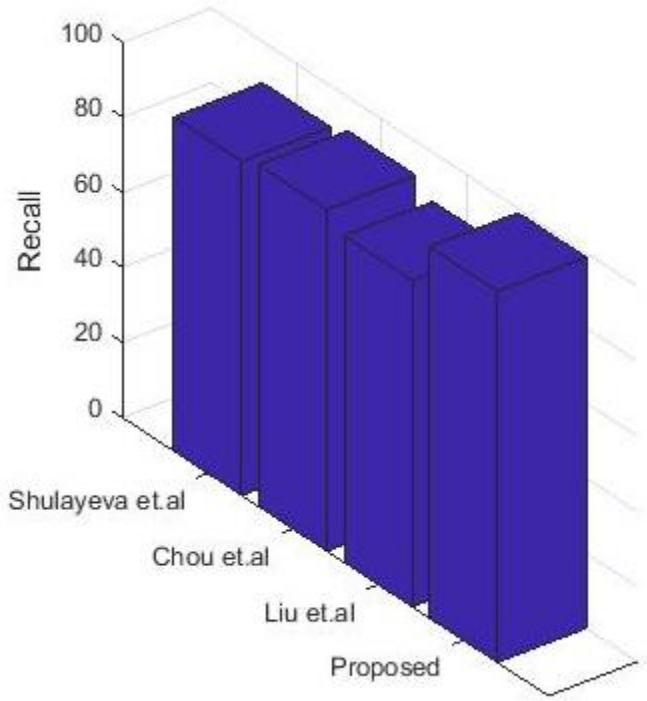

Fig. 10: Performance of Recall for Land record data.

Fig (8), (9), (10), shown as the Comparison of Accuracy, Precision, and Recall for Land Record Data in the existing system and the proposed system. Better performances of user need the retrieval information as accuracy, precision, and recall in the proposed system.

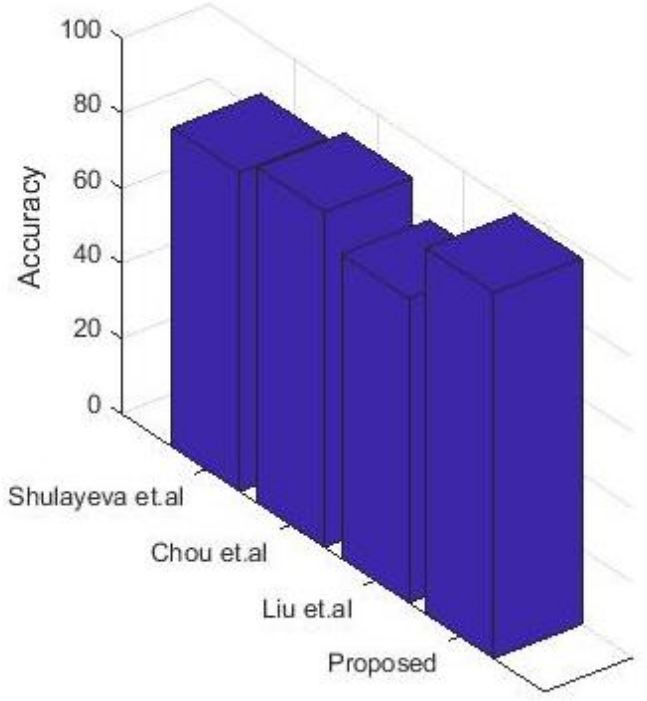

Fig. 11: Performance of Accuracy for Andhra Pradesh Data.

DOI https://doi.org/2024-JICS-22459 | Generated on 2025-04-15 10:49:18

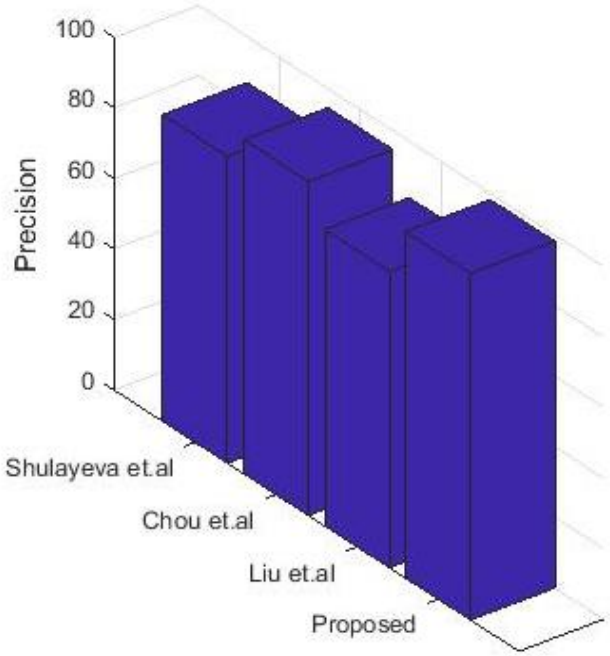

Fig. 12: Performance of Precision for Andhra Pradesh Data.

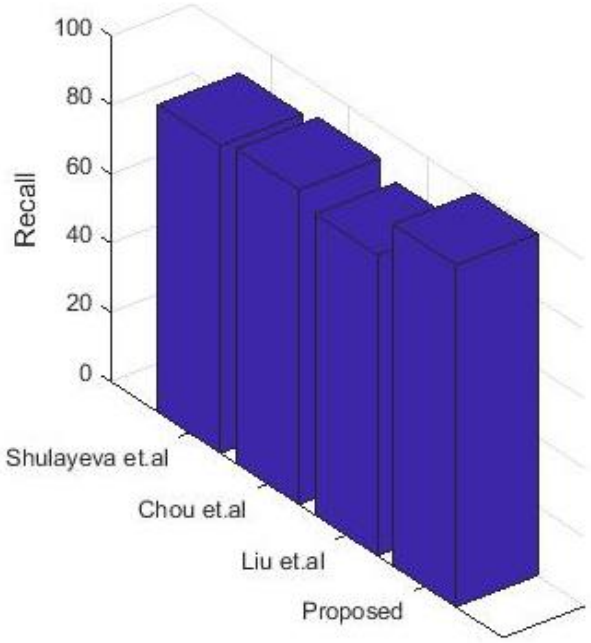

Fig. 13: Performance of Recall for Andhra Pradesh Data.

Fig (11), (12), (13) as shown the performances of Accuracy, Precision, and Recall for Andhra Pradesh Data in the existing system and the proposed system. The proposed system overlaps the retrieved information of accuracy, precision, and recall for the existing system.

5. Conclusions and future works

The main purpose of this paper is to provide easy access and retrieval of land-related information to the common user. To achieve the great accuracy in this article we apply some techniques. In our proposed work, the land information is gathered from resources then Artificial Neural Network classifies the information. Mongo DB stores all information and Mapping Co-ordinator helps to provide access rights and security. Finally, Optimized Fuzzy Based Ranking function based on Cat Swarm Optimization utilized to retrieve the relevant information. By using this framework the following benefits are achieved.

1. Mongo database can store the large quantity of data. It does not have any restriction or limitations.
2. This proposed work allows decision makers and policy planners to access relevant data (current data).
3. It improves the accuracy of the system.
4. ANN classifies helps to classify the content which stores the information in a continuous memory location. It can be performed in a clear technique to without any problem that applied in system modeling and system identification.
5. Mapping co-ordinator provides high security to the user.
6. In addition to that of user details in our proposed work is able to store graphical images.
7. Mapping co-ordinator is responsible for planning, implementing, coordinating and administration of citywide data.
8. It allows an authority to upload/download the new data.
9. It establishes the restriction on access to restricted information.
10. It provides great integrity assurance.

6. References

[1] March, Salvatore T., and Alan R. Hevner, Integrated decision support systems: A data warehousing perspective, Decision Support Systems 43, no. 3, pp. 1031-1043, 2007.

[2] Himanshu Gupta, and Inderpal Singh Mumick, Selection of views to materialize in a data warehouse, IEEE Transactions on Knowledge and Data Engineering 17, no. 1, pp. 24-43, 2006.

[3] Thilini Ariyachandra, and Hugh Watson, Key organizational factors in data warehouse architecture selection, Decision support systems 49, no. 2, pp. 200-212, 2010.

[4] Francisco Araque, Alberto Salguero, Maria M. Abad, Application of data warehouse and Decision Support System in soaring site recommendation, Information and Communication Technologies in Tourism, pp. 308-319, 2006.

[5] Haluk Demirkan, and Dursun Delen, Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud, Decision Support Systems 55, no. 1, pp. 412-421, 2013.

[6] Matthew Smith, Christian Szongott, Benjamin Henne, and Gabriele von Voigt, Big data privacy issues in public social media, 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST), pp. 1-6, IEEE, 2012.

[7] Samet Ayhan, Johnathan Pesce, Paul Comitz, David Sweet, Steve Bliesner, and Gary Gerberick, Predictive analytics with aviation big data, Integrated Communications, Navigation, and Surveillance Conference (ICNS), pp. 1-13, IEEE, 2013.

[8] Katharina Ebner, Thilo Bühnen, Nils Urbach, Think Big with Big Data: Identifying Suitable Big Data Strategies in Corporate Environments, IEEE, pp. 978-1-4799-2504, 2014.

[9] Bello-Orgaz, Gema, Jason J. Jung, and David Camacho, Social big data: Recent achievements and new challenges, Information Fusion 28, pp. 45-59, 2016.

[10] Torsten Priebe, Stefan Markus, Business Information Modeling: A Methodology for Data-Intensive Projects, Data Science and Big Data Governance, IEEE International Conference on Big Data (Big Data), 2015.

[11] Gang Pan, Guande Qi, Zhaohui Wu, Daqing Zhang, and Shijian Li, Land-use classification using taxi GPS traces, IEEE Transactions on Intelligent Transportation Systems 14, no. 1, pp. 113-123, 2013

[12] Vijay Katkar, Surupendu Prakash Gangopadhyay, Sagar Rathod, and Aakash Shetty, Sales forecasting using data warehouse and Naïve Bayesian classifier, Pervasive Computing (ICPC), 2015 International Conference on, pp. 1-6. IEEE, 2015.

[13] Jui-Sheng Chou, Min-Yuan Cheng, Yu-Wei Wu, and Anh-Duc Pham, Optimizing parameters of support vector machine using the fast messy genetic algorithm for dispute classification, Expert Systems with Applications 41, no. 8, pp. 3955-3964, 2014.

[14] Haluk Demirkan, and Dursun Delen, Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in the cloud, Decision Support Systems 55, no. 1, pp. 412-421, 2013.

[15] Torsten Priebe, Stefan Markus, Business Information Modeling: A Methodology for Data-Intensive Projects, Data Science and Big Data Governance, IEEE International Conference on Big Data (Big Data), 2015.

[16] Vikas S Shah, An Integrative Methodology to Enable Compliance aware Data Services Accelerating Consumer

Regulatory Governance Paradigms of an Enterprise, 13th IEEE Annual Consumer Communications & Networking Conference (CCNC), 2016.

[17] Osden Jokonya , Jan H. Kroeze, John A. van der Poll, A Framework to Assist Organizations with IT Adoption Governance, IEEE 6th International Conference on Cloud Computing Technology and Science, 2014.

[18] Veerendra Kumar Rai, Ashish Kumar Jha, Abhinay Puvvala, IT Service Management: A Typological Classification of Policies for Governance, IEEE 17th Conference on Business Informatics, 2015.

[19] Lukasz Golab, Theodore Johnson, and Vladislav Shkapenyuk, Scalable Scheduling of Updates in Streaming Data Warehouses, IEEE Transactions on Knowledge And Data Engineering, 2012.

[20] Mehdi Kashfi, Abdolreza Hajmoosaei, Optimal Distributed Data Warehouse System Architecture, IEEE Fourth International Conference on Big Data and Cloud Computing, 2014.

[21] Jinglan Zhang, Kai Huang, Mark Cottman-Fields, Anthony Truskinger, Paul Roe, Shufei Duan, Xueyan Dong, Michael Towsey, and Jason Wimmer, Managing and analysing big audio data for environmental monitoring, Computational Science and Engineering (CSE), IEEE 16th International Conference, pp. 997-1004, 2013.

[22] Bingwei Liu, Erik Blasch, Yu Chen, Dan Shen, and Genshe Chen. Scalable sentiment classification for big data analysis using Naive Bayes Classifier, IEEE International Conference on, pp. 99-104, 2013.

[23] Lisette García-Moya, Shahad Kudama, María José Aramburu, and Rafael Berlanga, Storing and analysing voice of the market data in the corporate data warehouse, Information Systems Frontiers 15, no. 3, pp. 331-349, 2013.

[24] Jemal H. Abawajy, Andrei Kelarev, and Morshed Chowdhury, Large iterative multitier ensemble classifiers for security of big data, IEEE Transactions on Emerging Topics in Computing 2, no. 3, pp. 352-363, 2014.

[25] Ahmad Taher Azar, and Aboul Ella Hassanien, Dimensionality reduction of medical big data using neural-fuzzy classifier, Soft computing 19, no. 4, pp. 1115-1127, 2015.

[26] Olga Shulayeva, Advaith Siddharthan, and Adam Wyner, Recognizing cited facts and principles in legal judgements, Artificial Intelligence and Law 25, no. 1, pp. 107-126, 2017.