

Evaluation of Data Mining Classification Algorithm by Monitoring Fall Detection Anomaly using Discriminative Approach

R. Poorni¹ and P. Madhavan²

¹School of Computing, SRM Institute of Science and Technology, Kattankulathur, Tamilnadu, India

Email: poorniram21@gmail.com

²Department of Computing Technologies, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Tamilnadu, India

Email: madhavap@srmist.edu.in

Abstract—Data mining plays a vital role in various sectors such as health care, finance, business, intrusion detection, scientific analysis and research. The user health behavior can be classified using various data mining models so that the analysis can be used for future predictions on the healthcare unit. The proposed model compares and evaluates various discriminative classification algorithms and shows the result of best classification model by applying it to the fall detection anomaly dataset.

Index Terms— Classification models, Data mining, Evaluation Techniques, Discriminative algorithms.

I. INTRODUCTION

Data mining being a subset of data science allows us to gain clear insight of the data to be analyzed and various decisions that can be attained. It transfers raw information into knowledge by providing various patterns to the input data. The huge volume of data is handled to extract the knowledge over the given domain. One such information gain from the given data is classification which classifies the object of similar category that contains the same behavior depending upon the various attributes or elements the object contains. This helps in organizing the voluminous data to improvise its quality. Data classification assign labels to all the data points from the predefined set of provided labels. Since the complete process depends on the predefined set of labels a detailed knowledge on the historical data is required. The classification algorithm analyzes the predefined labelled dataset to produce the result to the new data and thus it falls under supervised learning mechanism. The model is trained with a well labelled set of data and is used under the testing data to perform different set of tasks in data mining. The input value (x) is mapped to the function with the output variable (y) which is best suited as $y=fn(x)$. The suitability of the algorithm is based on the accuracy provided by various algorithms. The models can be generally categorized as discriminative and generative models. Discriminative models can be used under supervised learning algorithms such as classification or for regression. Estimating the class of the given data is done based on conditional probability. They define a clear boundary between labels used in the dataset and thus reduces the error rate and outliers hence provides more accuracy. The computational complexity of discriminative models are cheaper when compared to generative approach. The major drawback of working on

the discriminative model is that it requires a huge amount of data to produce the correct result and should not contain any missing data. The discriminative classifier model involves training the function estimation $f_n: x \rightarrow y$ or the probability of the cases $P(Y|X)$. The training data is analyzed to calculate the values of $P(Y|X)$. The popular models of discriminative approach includes Logistic regression, Support Vector Machine (SVMs), Traditional neural networks, Nearest neighbor, Conditional Random Fields (CRFs), Decision Trees and Random Forest. The proposed work discusses and compares the different discriminative approach models over fall detection dataset to find the best suit algorithm.

II. LITERATURE REVIEW

Syeda Farha Shazmeen, et al., [13] proposed a research work on the title “Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis” where various data mining applications and classification techniques are used and implemented with medical data set to check the efficiency of each algorithm. The performance of the algorithms are improved by performing preprocessing technique and selecting the appropriate features and classification is done by identifying new class labels. The attributes are selected based on the measure of attributes with high information gain which helps in least feature selection resulting in low computational complexity. The outcome shows that the neural network produced the highest performance for the given data set.

Shovan Chowdhury, Marco P. Schoen, [9] proposed a paper on “Research Paper Classification using Supervised Machine Learning Techniques” evaluating different machine learning techniques such as Support Vector Machines, Naive Bayes, K-Nearest Neighbor and Decision Tree for classifying published peer reviewed documents. It extracts the useful information from the given abstracts and classifying the publications into the following category as social, science and business. The parameters used for evaluation are precision, recall, F1 score and accuracy. For preprocessing the data, the process of tokenization in natural language processing techniques is used by dividing the paragraphs into useful words as tokens. It then removes unnecessary characters and change all the words to a same case to avoid ambiguity. The feature text is extracted by two types of vectorizer as Bag of Words and TF-IDF vectorizer. Then the data is given for the process of classification based on the aforementioned algorithms. The efficiency of these algorithms are compared and the outcome shows that SVM method produces the highest result whereas the decision tree dint work well for the text classification.

Michael P. Dessauer, Sumeet Dua, [7] proposed a paper on the title “Discriminative Features and Classification Methods for Accurate Classification” where feature selection strategies are developed. It gains the complete knowledge about the attributes or the elements in the data set and ranks the attributes to identify the minimal number of features that could provide the best output for the selected data. The factors selected for calculating the accuracy and efficiency are specificity and sensitivity. A large amount of ariel and land-based images are taken as input and automated classifiers are applied to the automated classifiers and the results are compared among several algorithms. The edge and moment descriptors are combined to produce high accuracy while performing classification and evaluation. The information knowledge includes wavelet transformation and statistics of the texture. The decision rules are developed to choose the adaptive algorithm based on conditions and data types. The levels of wavelet signals from the input image are approximated and the feature descriptors are identified that provides with the maximum separation of classification. The target descriptors are calculated based on the multi-scale wavelet representation and the best results are achieved from the level 1 and 2. For the textured data the edge histogram and invariant moments performed the best among various algorithms. Thus the combination of edge histogram and invariant moment are used at a wavelet level of 0 to 2 for the high resolution data.

Elvis Togbana, Djemel Zioua, [4] proposed a paper titled “Classification using Hierarchical mixture of Discriminative learners: How to achieve high scores with few resources?”. A generalized framework has been proposed over the hierarchical mixture of experts model. More than one discriminative models are combined using weights that are dependent over the given input. The gaussian weights and linear softmax are used as weights and a two level tree is generated. The generalized framework shows how the resources for classification can be reduced based on the weights. The minimal number of classifiers is used to classify the non-linear data. The major task is divided into several subtasks based on the weights provided. Two models were developed where model 1 uses different weighing function at each level and model 2 uses same weighing function throughout the model. This helps in dividing the tasks among classifiers and also reduces the number of classifiers used. The training time of model 2 is shorter but model 1 provided more accuracy when the number of classifiers are reduced.

Rajat Raina, Yirong Shen, Andrew Y. Ng, Andrew McCallum [8] proposed the paper titled “Classification with Hybrid Generative/Discriminative Models”. The paper describes the combination of generative and discriminative model in which the attributes are trained to achieve maximum efficiency. The high dimensional attributes are trained using generative model to improve generative likelihood and subset of the attributes are trained using discriminative model which increases conditional likelihood. The complexity bound is calculated in order to identify the best fitting discriminative parameters. The feature occurrence and the feature set size decides the number of training examples. The hybrid model provides better accuracy compared to a pure generative or a discriminative model and produces less error rate.

III. METHODOLOGY AND EXPERIMENTATION

The proposed model compares the various discriminative algorithm that are majorly used in supervised machine learning algorithms that apply statistical approach. The model mainly focuses on to differentiate or to appropriate choose between the generative or the discriminative approach for the classification algorithms. It is identified that the model proves to outperform the state of art models and the proof are given below. Here we have considered the anomaly detection dataset for the analysis.

Dataset: the dataset contains the coordinates of the images of people falling which considered to be anomaly. The classification of data depends on the time series of data with three coordinates X, Y and Z representing the positions of the sensors at each samples. The four sensors which are placed at person chest, ankles and waist are represented using the attributes 010-000-024-033, 010-000-030-096, 020-000-032-221 and 020-000-033-111. The decision label is given as 0 for normal and 1 for anomaly of person falling down.

The dataset consists of 8 attributes with 30,000 records, details of which are given in Table 1, below.

TABLE I: ATTRIBUTES OF THE DATASET

S.no	Attribute	Description
1	X	Position of sensor
2	Y	Position of sensor
3	Z	Position of sensor
4	010-000-024-033	Representations of sensor activity
5	010-000-030-096	Representations of sensor activity
6	020-000-032-221	Representations of sensor activity
7	020-000-033-111	Representations of sensor activity
8	Anomaly	Falling/Normal-living event

The correlation coefficient provides the linear associated value between two vectors. It provides the relationship between two values. The correlation coefficient is calculated as correct predictions rate will be obtained with high relationship between attributes. The more the correlation nearer to the probability of one there is a good linear relationship given between the variables.

A. Decision Tree

Even though the decision tree can be used for regression problems it is majorly used for the classification algorithms. It provides a tree structure by having the decision rules as branches which can usually have more than one branches, features of the dataset as internal nodes and the outcome of the method is represented as leaf node. For a given problem depending on a specific condition this provides all the conceivable solutions. It starts by checking the root node and the value in the dataset and based on the decision rule it chooses one internal node and repeats the process until it reaches the leaf node. To select the attributes for the node for implementing it in the decision tree the information gain parameter is used and given as the difference of the product of the weighted average of the entropy of each feature to the Entropy of the given samples.

Entropy provides the randomness or the uncertainty in the data and the value should be low thus provides a high information gain.

Table II and Figure 1 given below, shows the Decision Tree tested with respect to various percentage of data splitting and its corresponding correlation coefficient.

B. Random Forest

Random forest works by the combination of different classification models for problems that are complex and hence it improves the performance. It segregates the given dataset into various subset and contains the decision trees for these subsets. The accuracy of the prediction is identified by getting the average of the subset values. Having more number of subset will increase the accuracy of the model. The final prediction of the output model

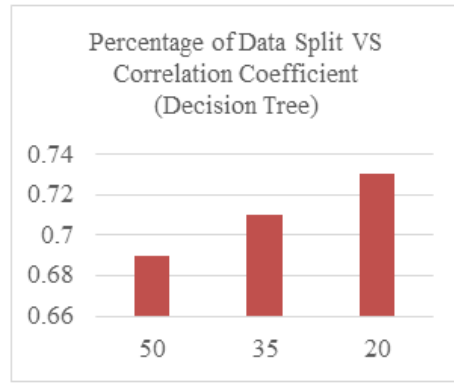


Fig. 1. Percentage of Data Split VS Correlation Coefficient for Table II

TABLE II: DECISION TREE DATA

PERCENTAGE SPLIT OF TESTING DATA	CORRELATION COEFFICIENT
50%	0.69
35%	0.71
20%	0.73

will be taken from the majority of the prediction that has happened for the subset. The correlation of the outputs of various decision trees must be low. This will avoid overfitting of data.

Table III and Fig. 2 given below, shows the Random Forest tested for different number of testing data and the correlation coefficient achieved.

TABLE III: RANDOM FOREST DATA

PERCENTAGE SPLIT OF TESTING DATA	CORRELATION COEFFICIENT
50%	0.61
35%	0.59
20%	0.68

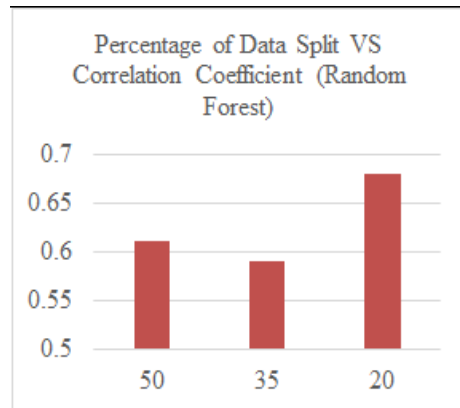


Fig. 2. Percentage of Data Split VS Correlation Coefficient for Random Forest model

C. K-Nearest Neighbour

The classification or prediction is done using the proximity of the data points. Based on the majority classes of the nearest values the classification of given data point is identified. The majority value is subject to the number of groups that are considered for the model and are iterative.

Table 5 and Figure 4 given below, shows the K-Nearest Neighbour tested for different number of testing data and the correlation coefficient achieved for the different values of dataset considered.

TABLE IV: K-NEAREST NEIGHBOUR DATA

PERCENTAGE SPLIT OF TESTING DATA	CORRELATION COEFFICIENT
50%	0.67
35%	0.67
20%	0.66

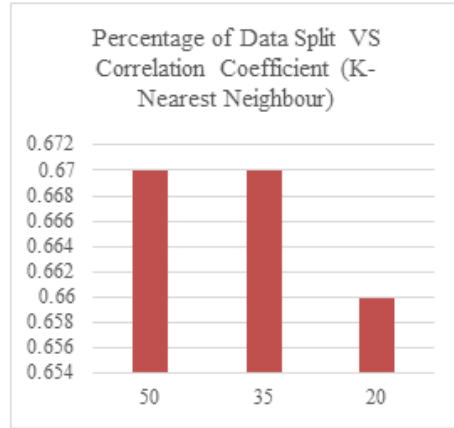


Fig. 4. Percentage of Data Split VS Correlation Coefficient for KNN model

IV. RESULTS AND EVALUATION

Figure 5 shows the graph for the accuracy of the above mentioned datamining techniques. After manipulating the dataset, it is identified that the decision tree performs well for the differential data and neural network works the least depending on the amount of dataset with which the model being trained.

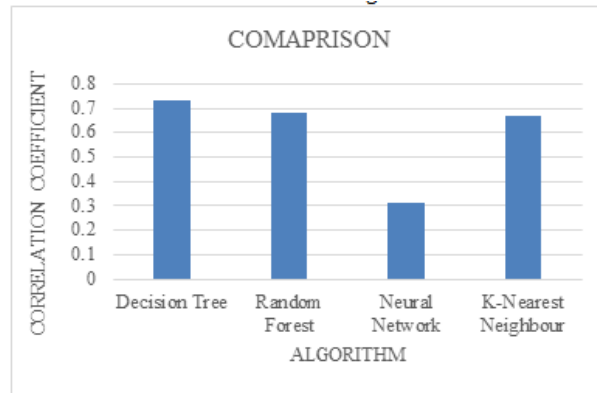


Fig. 5. Comparison of different datamining techniques

V. CONCLUSION AND FUTURE WORK

The proposed model compares some of the discriminative algorithms by taking a differential data and identified that more than generative model, discriminative model by a good amount of training data works well and concludes that decision tree performs well for the data and neural networks works better when the model is trained with a huge volume of data.

As a future work of the model, the attributes are selected and the model to be trained to make the model works well with the least amount of provided data as it reduces the cost and computational complexity.

REFERENCES

- [1] Shrestha and A. Mahmood, "Review of Deep Learning Algorithms and Architectures," in IEEE Access, vol. 7, pp. 53040-53065, 2019, doi: 10.1109/ACCESS.2019.2912200.

- [2] Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* 8, 53 (2021). <https://doi.org/10.1186/s40537-021-00444-8>
- [3] Bertino, Elisa & Nai Fovino, Igor & Provenza, Loredana. (2005). A Framework for Evaluating Privacy Preserving Data Mining Algorithms*. *Data Min. Knowl. Discov.* 11. 121-154. 10.1007/s10618-005-0006-6.
- [4] Elvis Togban, Djemel Ziou, Classification using Hierarchical mixture of Discriminative learners: How to achieve high scores with few resources?, *Expert Systems With Applications* (2017), doi: 10.1016/j.eswa.2017.11.046
- [5] Godbole, S., Sarawagi, S. (2004). Discriminative Methods for Multi-labeled Classification. In: Dai, H., Srikant, R., Zhang, C. (eds) *Advances in Knowledge Discovery and Data Mining, PAKDD 2004. Lecture Notes in Computer Science()*, vol 3056. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-24775-3_5
- [6] Lihua Luo, "Research on Image Classification Algorithm Based on Convolutional Neural Network", Published under licence by IOP Publishing Ltd *Journal of Physics: Conference Series*, Volume 2083, 2. Computational Science Citation Lihua Luo 2021 *J. Phys.: Conf. Ser.* 2083 032054 DOI 10.1088/1742-6596/2083/3/032054
- [7] Michael P. Dessauer and Sumeet Dua "Discriminative features and classification methods for accurate classification", *Proc. SPIE 7704, Evolutionary and Bio-Inspired Computation: Theory and Applications IV*, 77040L (15 April 2010); <https://doi.org/10.1117/12.853267>
- [8] Rajat Raina, Yirong Shen, Andrew Y. Ng, and Andrew McCallum. 2003. Classification with hybrid generative/discriminative models. In *Proceedings of the 16th International Conference on Neural Information Processing Systems (NIPS'03)*. MIT Press, Cambridge, MA, USA, 545–552.
- [9] S. Chowdhury and M. P. Schoen, "Research Paper Classification using Supervised Machine Learning Techniques," 2020 Intermountain Engineering, Technology and Computing (IETC), 2020, pp. 1-6, doi: 10.1109/IETC47856.2020.9249211.
- [10] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN COMPUT. SCI.* 2, 160 (2021). <https://doi.org/10.1007/s42979-021-00592-x>
- [11] Shantanu Godbole, Sunita Sarawagi, "Discriminative Methods for Multi-labeled Classification", *Advances in Knowledge Discovery and Data Mining*, 2004, Volume 3056, ISBN : 978-3-540-22064-0
- [12] Shazmeen, Syeda Farha, M. D. Mustafa and Ali Asadullah Baig. "Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis." *IOSR Journal of Computer Engineering* 10 (2013): 1-6.
- [13] Syeda Farha Shazmeen, Mirza Mustafa Ali Baig, M.Reena Pawar, "Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis" *IOSR Journal of Computer Engineering (IOSR-JCE)* e-ISSN: 2278-0661, p- ISSN: 2278-8727 Volume 10, Issue 6 (May. - Jun. 2013), PP 01-06 www.iosrjournals.org
- [14] Tarneem Elemam, Mohamed Elshrkawey, "A Highly Discriminative Hybrid Feature Selection Algorithm for Cancer Diagnosis", *The Scientific World Journal*, vol. 2022, Article ID 1056490, 15 pages, 2022. <https://doi.org/10.1155/2022/1056490>
- [15] Qing Lv, Suzhen Zhang, Yuechun Wang, "Deep Learning Model of Image Classification Using Machine Learning", *Advances in Multimedia*, vol. 2022, Article ID 3351256, 12 pages, 2022. <https://doi.org/10.1155/2022/3351256>.