Disease Prediction using K-means Clustering and Decision Tree

Asma Shaikh[1], Adwait Tapale[2], Shahid Sayyed[3], Mayur Dhage[4], Khandu Gunge[5]
[1,2,3,4,5]*MMCOE, Pune*

.***Abstract*****—**Classifiers can be either linear means Naive Bayes classifier or non-linear means decision trees. In this work we discuss with decision tree Naïve Bayes and k-means Classifiers can be either linear means Naive Bayes classifier or non-linear means decision trees. In this work we discusses with decision tree ,Naïve Bayes and k-means clustering .The Naive Bayes is based on conditional probabilities and affords fast, highly scalable model building and scoring. It scales linearly with the number of predictors and rows. And also build process is parallelized. Data Mining supports several algorithms that provide rules. Decision trees are among the best algorithms for data classification, providing good accuracy for many problems in relatively short time. Decision tree scoring is especially fast. The k-Means algorithm is a distance-based clustering algorithm that partitions the data into a predetermined number of clusters provided there are enough distinct cases.

Keywords**—** *Data Mining, Decision Tree, K-Means Algorithm*

I. INTRODUCTION

Data Mining is the extraction of implicit, previously unknown and rotationally useful information from data. Also, it is extraction of large database into useful data or information and that information is called knowledge. Data mining is always inserted in techniques for finding and describing structural patterns in data as a tool for helping that data and make prediction. Data mining consists of five major elements. First, extract, transform, and load transaction data onto the data warehouse system. Second, store and manage the data in a multidimensional database system. Third, provide data access to business analysts and IT professionals. Fourth, analyze the data by application software. Fifth, present the data in a useful format, such as a graph or table. Many data mining techniques are closely related to some of machine learning. Others are related to techniques that have been developed in statistics, sometimes called exploratory data analysis. We survey many techniques related to data mining and data classification techniques. We select clustering algorithm k-means to improve the training phase of Classification. Learning classification methods in data mining can be classified into three basic types: Supervised, unsupervised and reinforced.

1.1 Supervised Learning

In supervised learning, we are given a data set and already know what our correct output should look like, having the idea that there is a relationship between the input and the output. Supervised learning problems are further categorized into regression and classification problems.

1.2 Unsupervised Learning

On the contrary to Supervised learning, Unsupervised learning allows us to approach problems with little or no idea what our results should look like. We can derive structure from data where we don't necessarily know the effect of the variables. We can derive this structure by clustering the data based on relationships among the variables in the data. With Unsupervised learning there is no feedback based on the prediction results. For example, take a collection of 1,000,000 different genes, and find a way to automatically group these genes into groups that are somehow similar or related by different variables, such as lifespan, location, roles, and so on. This is a good example of clustering. Whereas, for a non-clustering problem such as "Cocktail Party Problem", it helps in identifying voices music from a mesh of sounds at a cocktail party.

1.3 Reinforcement Learning

Reinforcement Learning is, when exposed to an environment, how the machine train itself using trial and error. Machine mainly learns from past experiences and tries to perform best possible solution to a certain problem. In past couple of years, a lot of improvements in this particular area has been seen. Main example includes DeepMind's Alpha Go, beating the champion of the game Go in 2016.

Literature Survey

Author	Year	Data Mining Tool	Techniques used	Accuracy
Abhishek et al.	2013	WEKA 3.6.4	J48	95.56%
			Naïve Bayes	92.42%
			Neural Network	94.85%
Chaitrail et al.	2012	WEKA 3.6.6	Neural Network	100%
Nidhi et al.	2012	WEKA 3.6.6 TANAGRA .NET	Naïve Bayes	99.52%
			Decision Tree	52.33%
			Neural Network	96.5%
Vikas Chaurasia et al.	2013	WEKA	CART	83.49%
			ID3	72.93%
			Decision Table	82.50%
Hlaudi Daniel Masethe et al.	2014	WEKA	J48	99.074%
			REPTREE	99.74%
			Naïve Bayes	97.22%
			Bayes Net	98.14%
			Simple CART	99.74%
Rashedur et al.	2013	WEKA	Neural Network	79.19%
		TANAGRA	Fuzzy logic	83.85%

II. PROPOSED MODEL

2.1 K-Means Algorithm

K means algorithm is like a dividing based clustering algorithm, is to classify the given data objects into n different clusters over the iterative, converging to a local minimum. The results generated clusters are minimized and independent.

Input: C= {c1, c2, c3…..cn}, cluster sets,

D= {d1, d2, Dn} data sets

Output: find mean value µi

Begin

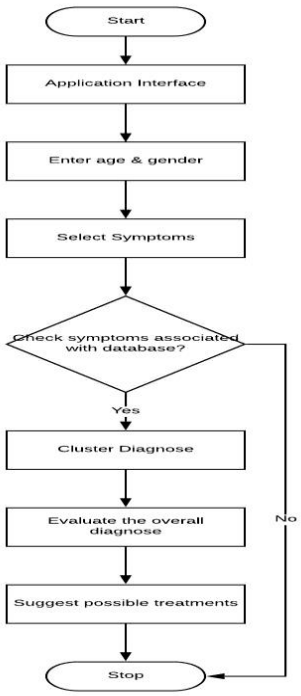
Choose any cluster from Data set D

Repeat While (Cj € D)

Assign Z as a Cluster centric

Select similar data Compute Mean Value µi

End



2.2 Decision Tree Algorithm

Decision trees are combined of computational and mathematical techniques to aid the representation, generalization and categorization of a given set of data. A Decision tree is a format which contains a root node, branches, and leaf nodes. Each internal node denoted as check on associate degree attribute, every branch denoted as the end result of a check and every leaf node denoted as a category label. The topmost node within the tree is called as root node. The main goal is to produce a model that predicts the value of a required variable based upon many input variables the decision tree model also uses the prediction-based rules classification. The known label of test data is compared along with the classified result. Accuracy rate is calculated based on the percentage of test set samples.

Algorithm for Decision Tree

Step 1: The leaflet is labeled with the same class if the instances belong to the same class.

Step 2: For each parameter, the potential information will be evaluated and the gain in information will be taken from the test on the parameter.

Step 3: Finally, the best parameter will be selected based on the present selection parameter.

Input: Attributes (a1, a2, a3…an)

Output: Predicted value Pv

Begin

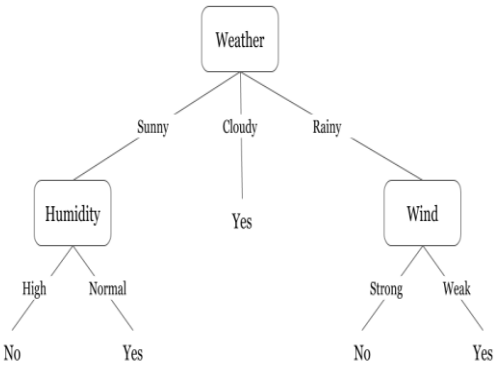
Where R – Root, B-Branches, Lf - Leaf nodes

Select each attribute (Aj)

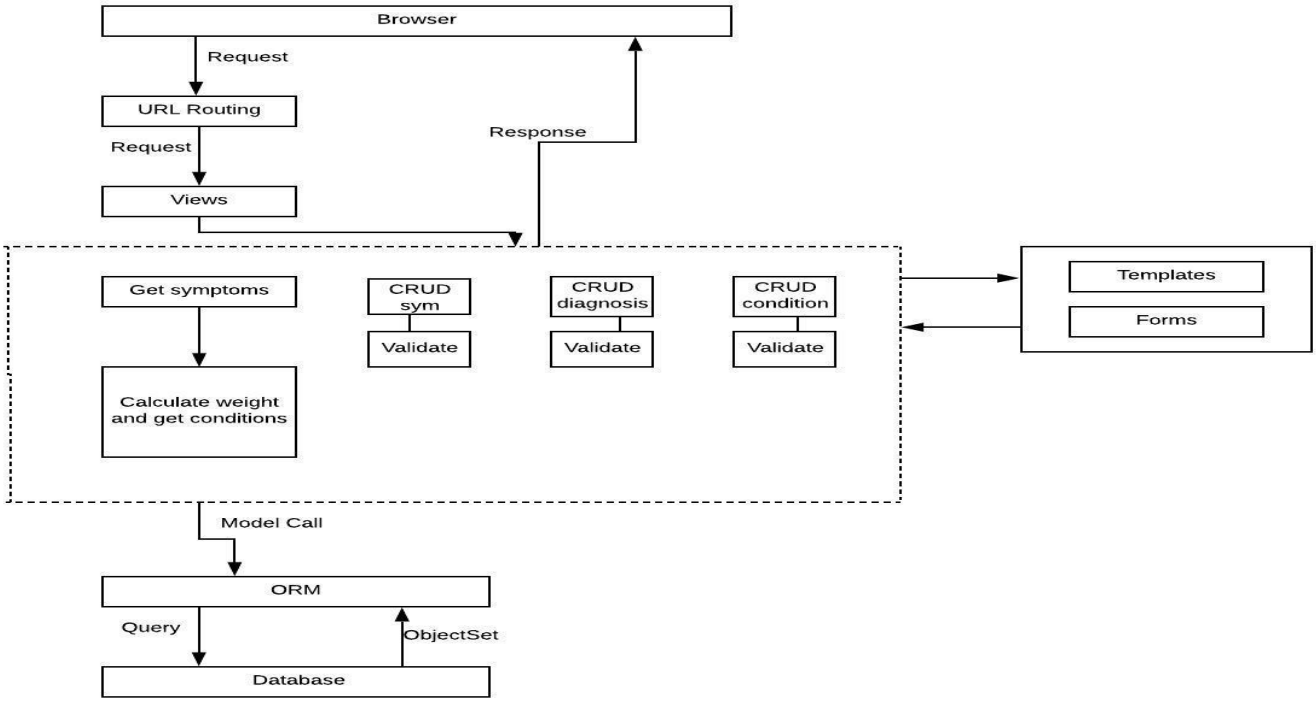
Calculate potential information Pi

Find best attribute based on the prediction

End



III. SYSTEM ARCHITECTURE



IV. CONCLUSION

As per the implementation of this project, data classification with decision tree is easy with compared to other method. Because of previous records and pictorial view, the task of categorization of data becomes easy. Once the result obtained, it can be reused for next research. This research depicts on compares reformulated decision tree with standard decision tree for dataset. Our comparison is from threshold (complexity) from low to high with reference to the testing accuracy. With this research a set of thresholds taken to show that our method gives better accuracy with decision tree rather than K- Means. The dataset is of type text and number. As the threshold is increased, the time taken to train decision tree is to be decreased. The advantage of decision tree is that it provides a theoretical framework for taking into account not only the experimental data to design an optimal classifier, but also a structural behavior for allowing better generalization capability. Almost every aspect of K-means has been modified Distance measures, Centroid and objective definitions, Overall process, Efficiency Enhancements, Initialization. In this paper we have presented a classification system that improvesclassification accuracy of any given decision tree algorithm by combining it with a clustering algorithm. The results exceeded our expectation, since clustering algorithms operate blindly (i.e. not taking the class into account) over the data, but yet manage to improve the accuracy of the system greatly, when compared to the basic system. In future research, more attention will be paid to studying cluster structure sensitivity that reflects reallocation of classes as well as to studying clustering quality. The K-means algorithm is a popular data-clustering algorithm. However, one of its drawbacks is the requirement for the number of clusters, to be specified before

The algorithm is applied. This paper first reviews existing methods for selecting the number of clusters for the algorithm.

Factors that affect this selection are then discussed and a new measure to assist the selection is proposed. The paper concludes with an analysis of the results of using the proposed measure to determine the number of clusters for the K means algorithm for different data sets.

REFERENCES

[1] Shekhar R. Gaddam, Vir V. Phoha, Kiran S. Balagani, "K-Means+ID3: A Novel Method for Supervised Anomaly Detection by Cascading KMeans Clustering and ID3 Decision Tree Learning Methods," Knowledge and Data Engineering, IEEE Transactions on , Vol. 19, No. 3, Pp. 345-354, March 2007.

[2] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu and Philip S. Yu, et al., "Top 10 algorithms in data mining", Knowledge and Information Systems, Vol. 14, No. 1, 1-37, DOI: 10.1007/s10115-007- 0114-2.

[3] Hang Yang, Fong, S, "Optimized very fast decision tree with balanced classification accuracy and compact tree size," Data Mining and Intelligent Information Technology Applications (ICMiA), 2011 3rd International Conference on, Pp.57-64, 24-26 Oct. 2011.

[4] Guang-Hua Chen; Zheng-Qun Wang; Zhen-Zhou Yu;, "Constructing Decision Tree by Integrating Multiple Information Metrics," Chinese Conference on Pattern Recognition, 2009. CCPR 2009, Pp.1-5, 4-6 Nov. 2009 DOI: 10.1109/CCPR.2009.5344133.

[5] Purusothaman G, Krishnakumari P. A Survey of Data Mining Techniques on Risk Prediction: Heart Disease. Indian Journal of Science and Technology. 2015 June; 8(12):1–5.

[6] A. K. Jain, M. N. Murty, P. J. Flynn, "Data clustering: a review", ACM Computing Surveys (CSUR) Surveys Homepage archive, Vol. 31, No. 3, 1999.

[7] Watanabe. N, "Fuzzy modeling by hyperbolic fuzzy k-means clustering," Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conferencr, Vol. 2, Pp.1528-1531, 2002 DOI: 10.1109/FUZZ.2002.1006733.

[8] Juanying Xie; Shuai Jiang; , "A Simple and Fast Algorithm for Global K-means Clustering," Education Technology and Computer Science (ETCS), 2010 Second International Workshop on , Vol. 2, Pp. 36-40, March 2010, DOI: