IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 1, Ver. V (Jan – Feb. 2015), PP 33-42
www.iosrjournals.org

Data mining Algorithm's Variant Analysis

[1]Kiranpreet, [2]Ramandeep Kaur, [3]Sourabh Sahota, [4]Prince Verma
[1]M.Tech CSE Dept., [2]M.Tech CSE Dept, [3]M.Tech CSE Dept, [4]Asst. professor CSE Dept.
CT Institute of Engg , Mgt.& Tech, Jalandhar (India)

***Abstract*:** *The Data Mining is extricating or mining information from extensive volume of information. Information mining frequently includes the examination of information put away in an information distribution center. This paper examines the three noteworthy information mining calculations: Association, classification and clustering. Classification is the procedure of looking connections between variables in gave database. Clustering comprises of anticipating a certain result focused around given information. Clustering is a procedure of dividing a set of data(or items) into a set of important sub-classes, called clusters. Execution of the 3 techniques is exhibited utilizing WEKA apparatus.*
Keyword Terms: *Data mining, Weka tools, Association, Classification and clustering algorithms.*

I. Introduction

Data mining discovers these examples and connections utilizing information investigation apparatuses and systems to assemble models. There are two sorts of models in information mining. One is prescient models i.e the methodology by which a model is made or decided to attempt to best anticipate the likelihood of a result.. An alternate is spellbinding models, is a scientific process that depicts certifiable occasions and the connections between elements in charge of them. The term Data Mining, otherwise called Knowledge Discovery in Databases (KDD) alludes to the nontrivial extraction of implied, conceivably valuable and already obscure data from information in databases [12]. There are a few information mining procedure are Association, Classification and Clustering [7]. Association principle learning is prominent and well hunt research system down revelation intriguing connection between variable in vast databases. Classification speaks to a managed learning strategy and additionally a measurable system for order. Classification calculations use distinctive procedures for discovering relations between the indicator characteristics' qualities and the focus on quality's qualities in the assemble information. Clustering is the undertaking of relegating a set of articles into gatherings (called groups) with the goal that the items in the same group are more comparable (in some sense or an alternate) to one another than to those in other clusters [13].

1.1 Data Mining Tasks

Data mining may involve six classes of tasks in common:
1. Anomaly detection – It is identification of unusual data records that might have errors and are un-interesting.
2. Association rule learning (Dependency modeling) – It is the process of searching relationships between variables in provided database.
3.Clustering – It is the task of discovering groups and structures in the data that can be similar in some way or another, without using any known structures.
4. Classification – It is the process of generalizing known structure to apply to new data.
5.Regression – It attempts to find a function that create model of data with the minimum errors.
6.Summarization – It is process of creating a compact representation of the data set, including report generation and visualization.

II. Data Mining Algorithms

Algorithm is set of operation, fact and rules. In data mining there is different type of algorithm that is:-

2.1 Association rule mining

The primary sub-issue can be further partitioned into two sub-issues: competitor itemsets generalition prepare and successive itemsets era process. We call those itemsets whose backing surpass the help edge as huge or regular itemsets, those itemsets that are normal or have the would like to be expansive or successive are called competitor itemsets. Different algorithms are proposed for ARM [1]:

2.1.1. AIS Algorithm [20]:

The AIS algorithm was the first calculation proposed for mining affiliation tenets. The calculation comprises of two stages. The main stage constitutes the era of the incessant item sets. The algorithm utilizes applicant era to catch the successive item sets. This is trailed by the era of the sure and successive affiliation leads in the second stage. The primary disadvantage of the AIS algorithm is that it makes numerous disregards

the database. Besides, it produces and numbers an excess of competitor item sets that end up being little, which obliges more space and squanders much exertion that ended up being futile.

2.1.2. APRIORI Algorithm [20,23]:

Apriori includes a stage for discovering examples called continuous item sets. A successive items et is a situated of things seeming together in various database records meeting a client detailed edge. Apriori utilizes a base up pursuit that lists each and every successive item set. This suggests with a specific end goal to deliver an incessant item set of length, it must create every last bit of its subsets since they excessively must be visit. This exponential multifaceted nature in a broad sense limits Apriori-like calculations to finding just short examples.

2.1.3. FP-TREE Algorithm [20,23]:

FP-tree-based algorithm is to parcel the first database to littler sub-databases by some part cells, and afterward to mine item sets in these sub-databases. Unless no new item sets can be discovered, the segment is recursively performed with the development of allotment cells. The FP-tree development takes precisely two sweeps of the exchange database. The primary output gathers the set of successive things, and the second sweep develops the FP-tree. Numerous different methodologies have been presented in the middle of with moment changes. At the same time fundamental among them and which are premise for new upcoming algorithm are Apriori and FP-tree Algorithm.

2.1.4. SETM Algorithm [23]:

SETM utilizes SQL to discover extensive thing sets. The calculation recalls Tides i.e. exchange Ids of the exchanges with the applicant thing sets. It utilizes this data rather than subset operation. This methodology has a detriment that if Ck needs to be sorted. Also additionally if Ck is so extensive there is no option fit in cradle assigned memory space, the circle is utilized as a part of FIFO methodology. At that point this obliges two outer sorts.

2.2 Classification

Classification anticipating a certain result focused around a given data. Order is an information mining capacity that appoints things in a gathering to target classes or classes. The objective of arrangement is to discover foresee the target class for each one case in the data. The algorithm tries to find connections between the characteristics that would make it conceivable to anticipate the result.

Classification is an vital information mining strategy with expansive applications. It is utilized to order everything in a set of information into one of predefined set of classes or gatherings. Arrangement calculation assumes a critical part in record order. In this examination, we have examined two classifiers specifically Bayesian and lazy.

Classification—A Two-Step Process

Model construction: Depicting a set of foreordained classes. each tuple/example is expected to fit in with a predefined class, as dictated by the class name attribute. The set of tuples utilized for model development: preparing set. The model is spoken to as grouping standards, choice trees or scientific equation.

Model usage: For classify future or obscure articles. Estimate accuracy of the model. The known name of test sample is contrasted and the ordered result from the model. Precision rate is the rate of test set examples that are accurately characterized by the model. Test set is free of preparing situated, generally over-fitting will happen.

2.2.1. Algorithm for Bayesian classification:- The Bayesian Classification speaks to a directed learning strategy and also a measurable system for order. Accept a hidden probabilistic model and it permits us to catch vulnerability about the model in a principled manner by deciding probabilities of the results. It can take care of symptomatic and prescient issues. Bayesian grouping gives commonsense learning algorithms and earlier information and watched information can be joined. Bayesian Classification gives a helpful viewpoint to understanding and assessing numerous learning calculations [13].

2.2.1.1 Bayesian Network[14,15]:

Bayesian system (BN) is additionally called conviction systems, is a graphical model for likelihood relationships among a set of variables gimmicks, This BN comprise of two segments. To begin with part is principally a directed acyclic chart (DAG) in which the hubs in the diagram are known as the irregular variables and the edges between the hubs or arbitrary variables speaks to the probabilistic dependencies among the relating irregular variables.second segment is a situated of parameters that portray the restrictive likelihood of every variable given its parents. bayesian system comprises of a structural model and a set of contingent

probabilities. The structural model is a coordinated diagram in which hubs speak to properties and circular segments speak to trait conditions. Property conditions are evaluated by contingent probabilities for every hub provided for its parents.

2.2.1.2. Naïve Bayes Algorithm [13]:

The Naïve Bayes Classifier procedure is focused around the alleged Bayesian theorem and is especially suited when the Trees dimensionality of the inputs is high. Regardless of its effortlessness, Naive Bayes can frequently outflank more modern characterization routines. Figure unequivocal probabilities for speculation, among the most down to earth methodologies to specific sorts of learning issues

2.2.1.3 Naïve Bayes Updatable [16]:

The name Naivebayes Updatable itself recommends that it is the updatable or enhanced form of Naivebayes. A default exactness utilized by this classifier when fabricate Classifier is called with zero training instances is of 0.1 for numeric properties and consequently it otherwise called incremental upgrade.

Naïve Bayes is better than Bayes network because

- Naïve Bayes is a quick execution speed algorithm.[17]
- The naive Bayes model is immensely engaging in view of its straightforwardness, style, and power. It is one of the most established formal classification algorithms, but even in its least difficult structure it is regularly shockingly successful.

Naive Bayes Classifier Introductory Overview: Naive Bayes Classifier procedure is focused around the alleged Bayesian theorem and is especially suited when the Trees dimensionality of the inputs is high. Regardless of its straightforwardness Naive Bayes can regularly outflank more modern order technique.

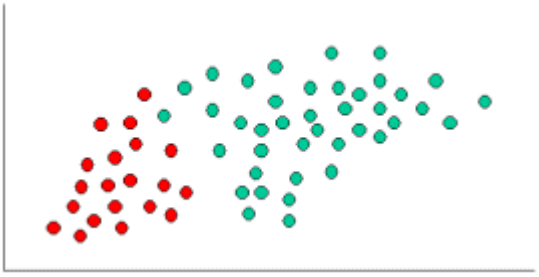

Fig No: 1

To show the idea of Naive Bayes Classification, consider the sample showed in the outline above. As indicated, the items can be named either GREEN or RED. Our errand is to arrange new cases as they arrive, i.e., choose to which class mark they have a place, based on the as of now leaving objects. Since there is twice the same number of GREEN protests as RED, it is sensible to accept that another case (which hasn't been watched yet) is twice as liable to have participation GREEN instead of RED. In the Bayesian examination, this conviction is known as the earlier likelihood. Earlier probabilities are focused around past experience, for this situation the rate of GREEN and RED items, and frequently used to anticipate results before they really happen.

Algorithm:

Given the obstinate example intricacy for learning Bayesian classifiers, we must search for approaches to decrease this many-sided quality. The Naive Bayes classifier does this by making a restrictive freedom presumption that drastically decreases the quantity of parameters to be assessed when demonstrating P(x|y), from our unique $2(2n - 1)$ to only $2n$.

Definition: Given arbitrary variables X,y and Z, we say X is restrictively free of Y given Z, if and just if the likelihood dissemination administering X is autonomous of the estimation of Y given Z; that is

$(\forall i, j, k)\ P(X=xi/Y=y j, Z=zk) = P(X=xi/Z=zk)$

Example:-

Consider three Boolean random variables to portray the current climate: Rain, Thunder and Lightning. We may sensibly attest that Thunder is independent of Rain given Lightning. Since we know Lightning reasons Thunder, once we know whether there is Lightning, no extra data about Thunder is given by the estimation of Rain. Obviously there is an agreeable reliance of Thunder is given by the estimation of Rain. Obviously there is an acceptable reliance of Thunder on Rain when all is said in done; however there is no restrictive reliance once we know the benefit of Lightning.

2.2.2. Algorithm for Lazy Classification:- Lazy learners store the training instances and do no genuine work until classification time. Lazy learning is a learning system in which speculation past the preparation information is deferred until a question is made to the framework where the framework tries to sum up the preparation information before getting inquiries.

2.2.2.1 IBL (Instance Based Learning):-

Instances based learning systems, for example, closest neighbor and mainly weighted relapse are adroitly clear methodologies to approximating genuine esteemed or discrete-esteemed target capacities. Example based techniques can likewise utilize more unpredictable, typical representations for occurrences.

2.2.2.2 IBK (K-Nearest Neighbor):-

IBK is a k-nearest neighbor classifier that uses the same separation metric. The quantity of closest neighbors can be defined expressly in the article editorial manager or decided naturally utilizing abandon one-out cross-approval center to a furthest utmost given by the tagged quality

2.2.2.3 K star:-

The K* algorithm can be characterized as a technique for cluster investigation which chiefly goes for the parcel of „n perception into „k clusters in which every perception fits in with the group with the nearest mean.

KNN is better than other lazy technique:

knn is straightforward and simple to actualize characterization system furthermore utilized for multi classes [19].

K-NN Algorithm Introduce:-K-NN is a lazy learning method focused around voting and separations of the k closest neighbors. K-Nearest Neighbor (KNN) calculation is a standout amongst the most well known learning calculations in information mining. The K-NN calculation for persistent esteemed target capacities Calculate the mean estimations of the k closest neighbors utilizing Euclidean distance.

Required three think:-

- The set of store record.
- Distance metric to compute the distance between record.
- The value of k, the number of nearest neighbors to retrieve.

2.3 Clustering

Cluster analysis was started in humanities by Driver and Kroeber in 1932 and acquainted with brain science by Zubin in 1938 and Robert Tryon in 1939and broadly utilized by Cattell starting as a part of 1943 for quality hypothesis arrangement in identity brain research. Cluster analysis[1] gatherings objects (perceptions, events)based on the data found in the information depicting the items or their connections. Cluster is a gathering of questions that have a place with the same class. The objective is that the items in a gathering will be comparable (or related) to one other and not quite the same as (or random to) the articles in different gatherings. Clustering is a methodology of apportioning a set of information (or articles) into a set of significant sub-classes, called clusters. A decent clustering strategy will create amazing clusters in which:

• the intra-class (that is, intra-bunch) likeness is high.

• the between class similitude is low.

The quality of a clustering result additionally relies on upon both the similitude measure utilized by the system and its usage. The quality of a clustering system is likewise measured by its capacity to find some or the majority of the shrouded examples. Then again, target assessment is hazardous: generally done by human/ master investing.

A. DBSCAN Clustering Algorithm

Density based clustering algorithm is one of the essential techniques for grouping in information mining. DBSCAN (for thickness based spatial clustering of uses with commotion) is an information clustering calculation proposed by Martin Ester, Hans-Peter Kriegel, Jorge Sander and Xiaowei Xu in 1996 The clusters which are structured focused around the thickness are straightforward and it doesn't restrict itself to the states of clusters. The density based gathering of clustering calculations speak to an information set in the same way as dividing methods; converting an example to a point utilizing the information traits of the source set. A standout amongst the most remarkable density based clustering calculations is the Dbscan[9]

DBSCAN separates data points into three classes:

 Core points: These are points that are at the interior of a cluster.

 Border points: A border point is a point that is not a core point, but it falls within the neighborhood of a core point.
 Noise points: A noise point is any point that is not a core point or a border point.
To discover a cluster, DBSCAN begins with a discretionary instance (p) in information set (D) and recovers all examples of D with admiration to Eps and Min Pts. The calculation makes utilization of a spatial information structure(r*tree) to place focuses inside Eps separation from the center purposes of the groups [2]. An alternate density based calculation OPTICS is presented in [3], which is an intuitive clustering calculation, lives up to expectations by making a requesting of the information set speaking to its density based clustering algorithm.It did clustering through becoming high thickness zone, and it can discover any shape if clustering(rong et al.,2004).

The thought of it was:
1. ε-neighbor: the neighbors in ε semi measurement of an article
2. kernal object: specific number (Minp) of neighbors in ε semi measurement
3. To an item set D, if object p is the ε-neighbor of q, and q is bit object, then p can get "immediate thickness reachable" from q.
4. To a ε, p can get "immediate thickness reachable" from q; D contains Mint objects; if an arrangementP1 ,p2 ,..., ,pn,p1=q, pn=p q , then i1 p can get "direct density reachable" from i p ,pi∈D≤ i≤n
5. To ε and MinP, if there exist a object o(o∈D) , p and q can get "direct density reachable" from o, p and q are density connected.

2.3.2. K-Mean Algorithm
The basic step of k-means clustering is simple. In the beginning we determine number of cluster K and we assume the centroid or center of these clusters. **K-means clustering** is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. Then the K means algorithm will do the three steps below until convergence Iterate until stable (= no object move group):
1. Determine the centroid coordinate
2. Determine the distance of each object to the centroids
3. Group the object based on minimum distance

2.3.2.1. Algorithmic steps for k-means clustering
Let $X = \{x_1,x_2,x_3,\ldots\ldots\ldots,x_n\}$ be the set of data points and $V = \{v_1,v_2,\ldots\ldots\ldots,v_c\}$ be the set of centers.
1) Randomly select 'c' cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..
4) Recalculate the new cluster center using:

$$v_i = (1/c_i)\sum_{j=1}^{c_i} x_i$$

where, 'c_i' represents the number of data points in i^{th} cluster.
5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3).
Example:-If our class (decision) attribute is tumor Type and its values are: malignant, benign, etc. - these will be the classes. They will be represented by cluster1, cluster2, etc. However, the class information is never provided to the algorithm. The class information can be used later on, to evaluate how accurately the algorithm classified the objects.

	Curvature	**Texture**	**Blood Consump**	**Tumor Type**
X1	0.8	1.2	A	Benign
X2	0.75	1.4	B	Benign
X3	0.23	0.4	D	Malignant
X4	0.23	0.5	D	Malignant

	Curvature	Texture	Blood Consump	Tumor Type
X1	0.8	1.2	A	Benign
X2	0.75	1.4	B	Benign
X3	0.23	0.4	D	Malignant
X4	0.23	0.5	D	Malignant

Fig No: 2

The way we do that, is by plotting the objects from the database into space. Each attribute is one dimension

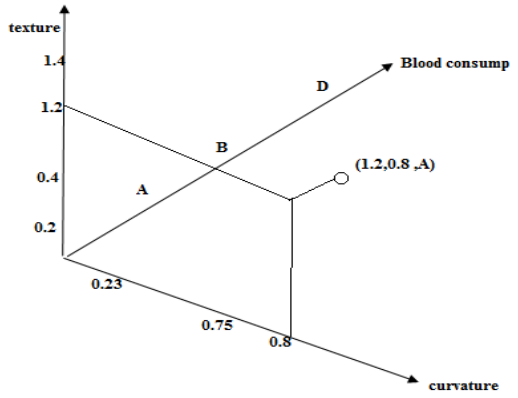

Fig No: 3

After all the objects are plotted, we will calculate the distance between them, and the ones that are close to each other – we will group them together, i.e. place them in the same cluster.

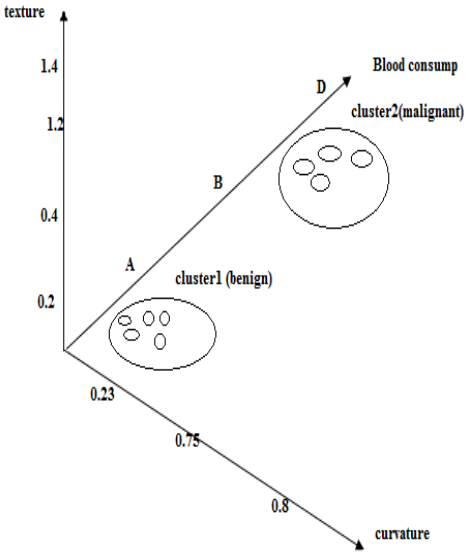

Fig No: 4

2.3.3 Hierarchical algorithm

Hierarchical clustering methods have attracted much attention by giving the user a maximum amount of flexibility. In data mining, **hierarchical clustering** (also called **hierarchical cluster analysis** or **HCA**) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

Agglomerative (bottom up)

1. Start with 1 point (singleton).
2. Recursively add two or more appropriate clusters.
3. Stop when k number of clusters is achieved.

Divisive (top down)

1. Start with a big cluster.
2. Recursively divides into smaller clusters.
3. Stop when k number of clusters is achieved.

In the general case, the complexity of agglomerative clustering is $O(n^3)$, which makes them too slow for large data sets. Divisive clustering with an exhaustive search is $O(n^2)$, which is even worse. However, for some special cases, optimal efficient agglomerative methods (of complexity $O(n^2)$) are known: SLINK for single-linkage and CLINK for complete-linkage clustering.

For example:-
suppose this data is to be clustered, and the Euclidean distance is the distance matric.Cutting the tree at a given height will give a partitioning clustering at a selected precision. In this example, cutting after the second row of the dendrogram will yield clusters {a} {b c} {d e} {f}. Cutting after the third row will yield clusters {a} {b c} {d e f}, which is a coarser clustering, with a smaller number but larger clusters.

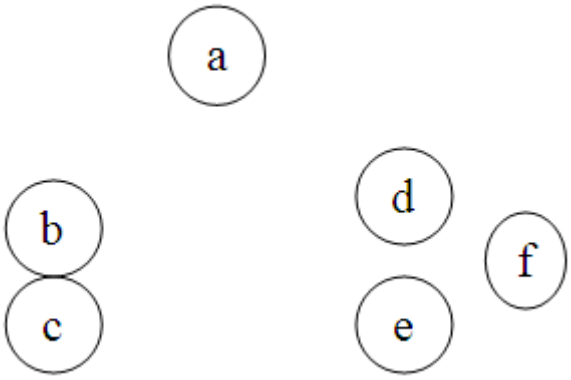

Fig No: 5

Raw data
The hierarchical clustering **dendrogram** (from Greek dendron "tree" and gramma "drawing") is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering. Dendrograms are often used in computational biology to illustrate the clustering of genes or samples. would be as such:

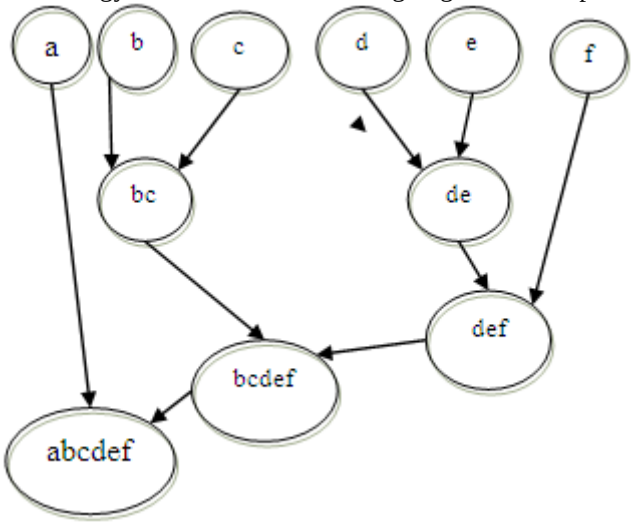

Fig No: 6

Traditional representation

This method builds the hierarchy from the individual elements by progressively merging clusters. In our example, we have six elements {a} {b} {c} {d} {e} and {f}. The first step is to determine which elements to merge in a cluster. Usually, we want to take the two closest elements, according to the chosen distance. Optionally, one can also construct a distance matrix at this stage, where the number in the i-th row j-th column is the distance between the i-th and j-th elements. Then, as clustering progresses, rows and columns are merged as the clusters are merged and the distances updated. This is a common way to implement this type of clustering, and has the benefit of caching distances between clusters. A simple agglomerative clustering algorithm is described in the single-linkage clustering page; it can easily be adapted to different types of linkage Suppose we have merged the two closest elements b and c, we now have the following clusters {a}, {b, c}, {d}, {e} and {f},

and want to merge them further. To do that, we need to take the distance between {a} and {b c}, and therefore define the distance between two clusters. Usually the distance between two clusters $\mathcal{A}$and $\mathcal{B}$is one of the following:

- The maximum distance between elements of each cluster (also called complete-linkage clustering):

Max{d(x,y): x□ A,y□ B}

- The minimum distance between elements of each cluster (also called single-linkage clustering):

Min{d(x,y): x□ A, y□ B}

- The mean distance between elements of each cluster (also called average linkage clustering, used e.g.UPGMA):

$$\frac{1}{|A|.|B|}\sum_{x\varepsilon A}\sum_{x\varepsilon B} d(x,y)$$

- The sum of all intra-cluster variance.

III. Result Using Weka Tool:-

WEKA

In 1993, the University of Waikato in New Zealand started development of the original version of Weka (which became a mixture of TCL/TK, C, and Make files). **Weka** (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand.

Dataset Use:-

1. Monk
2.Titanic
3. iris_discretized
4.CPU
5. lymphography

3.1. Result of apriori Algorithm with monk dataset

Table 1: Monk Problem Dataset Results.

Fitting Contact Lenses Dataset	**Apriori**			
	min sup=0.06	**min sup=0.07**	**min sup=0.08**	**min sup=0.09**
Confmin	0.0666	0.1333	0.1333	0.1333
Confmax	1	1	1	1
Confavg	0.35458	0.42041	0.42041	0.42041

Table 2: No of Rules (Monk Problem Dataset)

Fitting Contact Lenses Dataset	**Apriori**			
	min sup=0.06	**min sup=0.07**	**min sup=0.08**	**min sup=0.09**
Total Rules	2682	966	192	172

Table 3: Time Required in (Milliseconds) Monk Problem Dataset

Fitting Contact Lenses Dataset	**Apriori**			
	min sup=0.06	**min sup=0.07**	**min sup=0.08**	**min sup=0.09**
Time Required (milliseconds)	16290±100	2735±50	2735±50	2735±50

3.2 Result of K-NN and Naïve Bayes

Table 4: Evaluation of Naïve Bayes and Lazy Classifiers with Titanic Dataset

Algorithm	**Correctly Instance (%)**	**Incorrectly Instance (%)**	**Accuracy**			
			TP Rate (%)	**Recall (%)**	**Precision (%)**	**F-measure (%)**
Naïve Bayes	77.8283	22.1717	77.8	77.8	77.2	76.4
K-NN	79.055	20.945	79.1	79.1	82.1	75.9

Table 5: Evaluation of Naïve Bayes and Lazy Classifiers with Iris Dataset

Algorithm	Correctly Instance (%)	Incorrectly Instance (%)	Accuracy			
			TP Rate (%)	Recall (%)	Precision (%)	F-measure (%)
Naïve Bayes	94.6667	5.3333	94.7	94.7	94.7	94.7
K-NN	96.6667	3.3333	96.7	96.7	96.7	96.7

3.3. Result of k mean ,Hierarchal And density Based algorithms.

Table 6: Comparison result of algorithms using CPU dataset

Algorithm	No. of cluster	Cluster Instances	No. of Iterations	Sum of squared errors	Time taken to build model	Log Likelihood
K-Mean	2	0 :155(74%) 1:54(26%)	8	182.0143	0.02 sec.	
Hierarchical clustering	2	0:1(0%) 1:208(100%)	8		0.03 sec.	
Density Based clustering	2	0:151(72%) 1:58(28%)	8	182.0143	0.11 sec.	-94.74109

Table 7:Comparison result of algorithms using Lymphography dataset

Algorithm	No. of cluster	Cluster Instances	No. of Iterations	Sum of squared errors	Time taken to build model	Log Likelihood
K-Mean	2	0 :75(51%) 1:73(49%)	4	795.809	0.01 sec.	
Hierarchical clustering	2	0:146(99%) 1:2(1%)	4		0.13 sec.	
Density Based clustering	2	0:75(51%) 1:73(49%)	4	795.809	0.8 sec.	-15.33

IV. Conclusion

In this paper we conclude that in association rule mining apriori algorithm is best among other algorithms of association because Apriori involve frequent item set for candidate generation using bottom up search which requires producing all of its frequent subsets. In classification algorithm K-NN is best algorithm to other classification algorithms because k-NN is based on nearest neighbor and fast executes technique as compare Bayesian. And in clustering algorithm k-mean is better than other algorithms because k-mean is a simple and fast process it is easy to implement and it take less time to execute.

References

[1]. Yuni Xia, Bowei Xi ―Conceptual Clustering Categorical Data with Uncertainty Indiana University – Purdue University Indianapolis Indianapolis, IN 46202, USA.

[2]. Xu R. Survey of clustering algorithms.IEEE Trans. Neural Networks 2005.

[3]. B¨OHM, C., KAILING, K., KRIEGEL, H.-P., AND KR¨OGER, P. 2004. Density connected clustering with local Subspace preferences.In Proceedings of the 4th International Conference on Data Mining (ICDM).

[4]. R. Ng and J. Han. "Efficient and effective clustering method for spatial data mining". In: Proceedings of the 20th VLDB Conference,pages 144-155, Santiago, Chile, 1994.

[5]. Fei Shao, Yanjiao Cao ―A New Real-time Clustering Algorithm Department of Computer Science and Technology, Chongqing University of Technology Chongqing 400050, China.

[6]. Slava Kisilevich, Florian Mansmann, Daniel Keim ―P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos, University of Konstanz.

[7]. V. Filkov and S. kiena. Integrating microarray data by consensus clustering. International Journal on Artificial Intelligence Tools, 13(4):863–880, 2004.

[8]. [8] M. and Heckerman, D. (February, 1998). An experimental comparison of several clustering and intialization methods.Technical Report MSRTR-98-06, Microsoft Research, Redmond,WA.

[9]. Timonthy C. Havens. "Clustering in relational data and ontologies" July 2010.

[10]. Narendra Sharma 1, Aman Bajpai 2, Mr. Ratnesh Litoriya 3." Comparison the various clustering algorithms of weka tools" (ISSN 2250-2459, Volume 2, Issue 5, May 2012)

[11]. Bharat Chaudhari1, Manan Parikh2, KITRC KALOL" A Comparative Study of clustering algorithms Using weka tools. Volume 1, Issue 2, October 2012

[12]. J. Han and M. Kamber, (2000) "Data Mining: Concepts and Techniques," Morgan Kaufmann.

[13]. Vijaykumar,Vikramkumar,Trilochan " Bayes and Naive-Bayes Classifier".

[14]. Liangxiao Jiang, Harry Zhang, and Zhihua Cai "A Novel Bayes Model: Hidden Naive Bayes" ieee transaction on knowledge and data engineering,vol.21,no.10,October 2009.

[15]. Delveen Luqman Abd AL-Nabi, Shereen Shukri Ahmed "Survey on Classification Algorithms for Data Mining:(Comparison and Evaluation)" Computer Engineering and Intelligent Systems Vol.4, No.8, 2013.

[16]. V. Vaithiyanathan,K. Rajeswari, Kapil Tajane, Rahul Pitale "Comparison of different classification techniques using different dataset" International Journal of Advances in Engineering & Technology, May 2013. ©IJAET

[17]. XindongWu· Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang ·Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu ·Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg "Top 10 algorithms in data mining" Springer-Verlag London Limited 2007, 4 December 2007.

[18]. Hung-Ju Huang and Chun-Nan Hsu, Member, IEEE "Bayesian Classification for Data From the Same Unknown Class" IEEE Transactions On System , Man and Cybernetics-part B: Cybernetics, Vol.32, N0.2, April2002.

[19]. C. Lakshmi Devasena "Classification Of Multivariate Data Sets Without Missing Values Using Memory Based Classifiers-An Effectiveness Evaluation" International Journal of Artificial Intelligence & Applications (IJAIA), Vol.4,No.1,January2013.

[20]. R. Agrawal, T. Imielinski, and A. N. Swami, 1993. Mining association rules between sets of items in large databases. ACM SIGMOD International Conference on Management of Data, Washington, D.C., pp 207-216.

[21]. Jie Gao, Shaojun Li, Feng Qian "Optimization on Association Rules Apriori Algorithm" IEEE Conference, vol 2, pp 5901-5905, 2006.

[22]. Dongme Sun, Shaohua Teng, Wei Zhang, Haibin Zhu "An Algorithm to Improve the Effectiveness of Apriori" IEEE Conference, pp 385-390, Aug 2007.

[23]. R.Divya and S.Vinod kumar "Survey On AIS, Apriori and FP-Tree algorithms" International Journal of Computer Science and Management Research vol 1, issue 2, pp 194-200, September 2012..

[24]. Aastha Joshi, Rajneet Kaur,2013. Comparative Study of Various Clustering Techniques in Data Mining, International Journal of Advanced Research in Computer Science and Software Engineering(ISSN: 2277 128X, Volume 3, Issue .

[25]. Sang Jun Lee, Keng Siau "A review of data mining techniques" Industrial Management and Data Systems, University of Nebraska-Lincoln Press,USA, pp 41-46,2001

[26]. Peter Robb, Carlos Coronel. Database Systems: Design, Implementation and management. Cengage Learning, 8th Edition, 2009.

[27]. M.S. Chen, J. Han, and P.Yu "Data mining: an overview from a database perspective'', IEEE Transactions on Knowledge and Data Engineering, vol. 8, no. 6, pp. 866-883,1996

[28]. U. Fayyad, S.G.Djorgovski and N.Weir "Automating the analysis and cataloging of sky surveys'', Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, MA, pp. 471-94, 1996.

[29]. JaiWei Han ,Jian Pei ,Yiwen Yin & Runying Mao "Mining frequent patterns without candidate generation: A Frequent pattern tree approach" Data mining and knowledge discovery ,Netherlands, pp 53-87, 2004.

[30]. U. Fayyad, G. Piatetsky-Shapiro and P.Smyth ``From data mining to knowledge discovery: an overview'', Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, MA, 1996.

[31]. M.S. Chen, J. Han, and P.Yu "Data mining: an overview from a database perspective'', IEEE Transactions on Knowledge and Data Engineering, vol. 8, no. 6, pp. 866-883,1996.

[32]. Sang Jun Lee, Keng Siau "A review of data mining techniques" Industrial Management and Data Systems, University of Nebraska-Lincoln Press,USA, pp 41-46,2001Peter Robb, Carlos Coronel. Database Systems: Design, Implementation and management. Cengage Learning, 8th Edition,2009.

[33]. G.Kesavaraj And Dr.S.Sukumaran "A Comparison Study on Performance Analysis of Data Mining Algorithms in Classification of Local Area News Dataset using WEKA Tool" International Journal Of Engineering Sciences &Research Technology 2(10),October 2013.

[34]. Aaditya Desai And Dr. Sunil Rai "Analysis of Machine Learning Algorithms using WEKA" International Conference & Workshop on Recent Trends in Technology, (TCET) 2012 Proceedings published in International Journal of Computer Applications® (IJCA).

[35]. M. Kantardzic, Data Mining - Concepts, Models, Methods, and Algorithms, IEEE Press, Wiley-Interscience, 2003, ISBN 0-471-22852-4.

[36]. Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Elsevier 2006, ISBN 1558609016.