An Overview of Efficient Data Mining Techniques

Sandeep Dhawan
Director of Technology IT Department
OTTE New York
USA

Abstract:- **Data mining is the process of discovering associations within huge data set, finding data patterns, anomalies, changes and significant statistical structures in the data. Conventional data analysis techniques involve formulating a hypothesis and then validating it against the dataset. On the other hand data mining techniques automatically detect significant patters in the data and these patterns can be used to formulate algorithms. An important consideration in mining huge data sets is that the result or the pattern identified should be valid, understandable, useful and novel [1]. Not to go without saying that data warehousing and maintaining large databases also principally rely on the efficiency of robust, intelligent and at times novel data mining techniques. Today data mining (techniques) are employed in nearly every sector of corporate industry. From music industry to films maintenance, medicine to sports there's hardly any field of life without an input and integration of these data mining techniques. This paper focuses on presenting an overview of some of the most commonly used data mining techniques along with their applications. Techniques presented in this paper include sequence mining, clustering, classification, K nearest neighbors and association rule mining. Additionally, there's a sample example in each case to help understand the basic working of each technique. Underlying branches, algorithms and process for each of these techniques are also given. Pseudo code for algorithms is also mentioned where required to ensure readers understanding with respective graphs. Paper also gives a brief overview some of the pre and post-processing data mining techniques.**

Key Words: Data mining, KNN, Clustering, Classification, Association rule mining, Sequence mining.

1. DATA MINING PROCESS

Data mining is a modular process and it is completed in multiple stages. This section throws light upon the stages involved in carrying out a data mining research on a dataset (2, 3).

1.1. Explore Data Domain

Before dwelling into the algorithm designing and data accumulation process, profound understanding of application domain, in which research has to be carried out, must be developed. Having grasp over the application domain helps in accumulating better data sets and deciding what data mining technique should be applied to achieve the expected results.

1.2. Collect Data

All of the data mining algorithms are implemented on some data sets. Therefore a dataset sufficient enough to satisfy all the requirements of the algorithm being implemented must be gathered.

1.3. Refine and Transform the data

Datasets contain noise, outliners, missing values and other inconsistencies which need to be removed before data can be further processed for analysis and pattern extraction. A research carried out by Swine MHC (4) studied 173 data records and found that 36 of them were faulty or erroneous. Therefore it is extremely important to refine and transform data. (5)

1.4. Feature Selection

Dataset might consist of thousands of features, however for a particular problem, only a handful of features are required. Therefore, after a refined dataset has been obtain which is not contaminated by inconsistencies and noise, relevant features are selected to apply further processing. Feature selection is done via techniques like principal component analysis [7], Wilcoxson rank sum test [6], entropy analysis and Fisher Criterion [8].

1.5. Apply Relevant Algorithm

After data has been acquired, cleaned and features have been selected, algorithm is selected that will process the data and produce results [1]. Some of the most commonly used algorithms are clustering, association rule mining, decision tree, sequence mining etc. The details of these algorithms have been explained in the later sections.

1.6. Observe, Analyze and Evaluate

The last and final stage of data mining process is to observe, evaluate and find patterns in the results produced by the algorithm. Final conclusion is made on the basis of these evaluations [1].

2. DATA MINING TECHNIQUES

There are many data mining techniques currently in use by data scientists and they differ depending upon their efficiency, precision and the type of data upon which they operate [2]. Here however, we will only discuss 5 of them. Association rule mining, Classification, Clustering, KNN and Sequence mining.

3. ASSOCIATION RULE MINING [11][12]

Proposed by Agarwal et al in 1993[11], today it is extensively used by data mining fraternity. It assumes data to be categorical, hence not popular option for numerical data analysis. Initially it was used in 'Market Basket Analysis',[12] an analysis to find out relations between customer's shopped items. In explanation below, we will use the same example with association rule mining detailed description.

3.1. Example Case: Market Basket Analysis.

3.1.1. Data:

ConsiderItems I: {i_1, $i_{2,}$ ….,i_x}: A set of articles/items in the store i.e. Butter.
Transaction *t*: *t* is a set of items brought by a particular person, and $t \subseteq I$.
Transaction Database *T*: a set of transactions from stores database, $T = \{t_1, t_2, \ldots, t_n\}$.
Hence a transaction would be something like:
t1: {bread, butter, milk} t2: {apple, eggs, salt, sugar}
… tn: {biscuit, eggs, milk}

3.1.2. Rules:

I. A transaction *t* from transaction database T, contains *X*, a set of items (item set) in *I*, if $X \subseteq t$.

II. An association rule is a simple manifestation of the form: $X \rightarrow Y$, where *X*, Y $\subset I$, and $X \cap Y = \varnothing$

III. An item set is a set of items.E.g., X = {milk, butter, cereal} is an item set.

IV. A *k*-item set is an item set with *k* items.E.g., {milk, bread, cereal} is a 3-itemset and {eggs, butter} is a 2-itemset.

3.1.3. Measuring Standards:

Support (sup): The rule exists with support *sup* in *T* (the transaction data set) if sup percentage % of transactionscontain $X \cup Y$.

$$\text{I.e. } sup = Pr(X \cup Y).$$

Confidence (conf): The rule exists in T (the transaction data set) with conf, if conf percentage % of transactions that contain X also contain Y.

$$\text{I.e. } conf = \Pr(Y \mid X).$$

Hence an association rule is a simple pattern that states whenever *X* occurs, *Y* occur with a certain probability.

3.1.4. Measuring Support & Confidence:

Support count: X.count (denotation of support count) of an item set X, in a dataset T is the number of transactions in T that contains X. Assuming there are n transactions in T, then:

$$support = \frac{(X \cup Y).count}{n}$$

Confidence Count: In a dataset is the number of transactions in T that contains X against total no of items in item set X. i.e.

$$confidence = \frac{(X \cup Y).count}{X.count}$$

Table -1: Example of Support Measure.

TID	Items	Support=Occurrence/Total Support
1	*Chicken, Milk*	*Total Support=5*
2	*Beef, Cheese*	*Support[Beef, Cheese]=2/5=40%*
3	*Chicken, Clothes, Cheese, Milk*	*Support[Chicken, Milk]=3/5=60%*
4	*Beef, Chicken, Cheese*	*Support[Beef, Chicken, Cheese]=1/5=20%*
5	*Chicken, Milk*	

Table -2: Example of Confidence Measure.

TID	Items	Given X->Y: Confidence= Occurrence[Y]/ Occurrence[X]
1	*Apple, Butter, Cheese*	
2	*Apple, Butter, Detergents*	*Confidence {Apple->Butter}= 2/3=66%*
3	*Butter, Cheese*	*Confidence {Butter->Cheese}=3/4=75%*
4	*Apple, Cheese*	*Confidence {Apple, Butter->Cheese}=1/2=50%*
5	*Butter, Cheese, Detergents*	

3.2. Summary Association Rule Mining:

Currently there are many algorithms those implement Association based mining, however Apriori Algorithm (a two-step iterative process) is most largely associated with Association mining [12]. Another popular Algorithm is DLG Algorithm. Important to note here is that space of all association rules is exponential, O (2n), where n is the number of items in I. Additionally, associative mining exploits sparseness of data, and high minimum support (minsup) and high minimum confidence (minconf) values [12].

4. SEQUENCE MINING[13][18]

Sequence mining is closely related to associative rule mining [2]. However the principal difference between the two lies in the input dataset. Each row in the input dataset acts as a single data sample/data item for associative mining (as in transactions $t_1, t_2, ..t_n$ above), whereas a data sample (aka sequence in sequence mining) is split across multiple consecutive rows in input data, each row representing only one event, based on some predefined criteria. Hence each event acts as a transaction.
Sequential mining techniques are used for medical treatments, natural disasters, stocks and markets, DNA sequencing and gene structures [13].

4.1. Related Terms:

I. **Subsequence vs. Super sequence:**

Generally, a sequence is an ordered list of events, denoted as $< e_1\ e_2 \ldots e_l >$. Given two sequences $A=< a_1\ a_2 \ldots a_n >$ and $B=< b_1\ b_2 \ldots b_m >$. A is called a subsequence of B, denoted as A $\subseteq \beta$, if there exist integers $1 \leq j_1 < j_2 < \ldots < j_n \leq m$ such that $a_1 \subseteq b_{j1}$, $a_2 \subseteq b_{j2}, \ldots, a_n \subseteq b_{jn.}$
In which case, B is a super sequence of A:
E.g. A=< (ab), d> and B=< (abc), (de)>.

4.2. Sequential Pattern Mining(Sample Case):

Given a set of sequences (Activities) with a support threshold, we can find the complete set of frequent subsequences:

Sequence Database: Given a dataset, a sequence : < (ef) (ab) (df) c b >, An element may contain a set of items. Items within an element are unordered and we list them alphabetically. Hence <a(bc)dc> is a subsequence of <a(abc)(ac)d(cf)>.

Given support threshold min_sup=2, <(ab)c> is a sequential pattern.

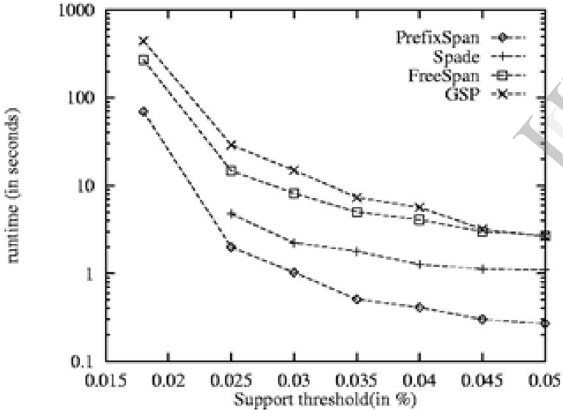
SID	Sequence
10	*<(ad)c(bc)(ae)>*
20	*<a(abc)(ac)d(cf)>*
30	*<eg(af)cbc>*
40	*<(ef)(ab)(df)cb>*

Table -3: Sample Sequence Data.[13]

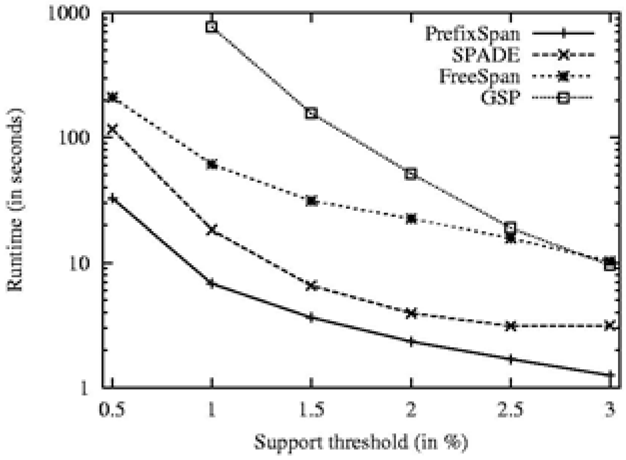
4.3. Common methods of Sequence based mining:

I. Apriori-based Approaches: GSP, SPADE etc.

II. Pattern-Growth-based Approaches: FreeSpan, PrefixSpan etc.



Graph -1: Performance Comparison of Approacheson Data Set C10T8S8I8.[13]



Graph -2: Performance Comparison of Approacheson Data Set**Gaelle**.-[13]

Here we will only discuss GSP, given its high performance evident from the chart.

4.4. GSP:

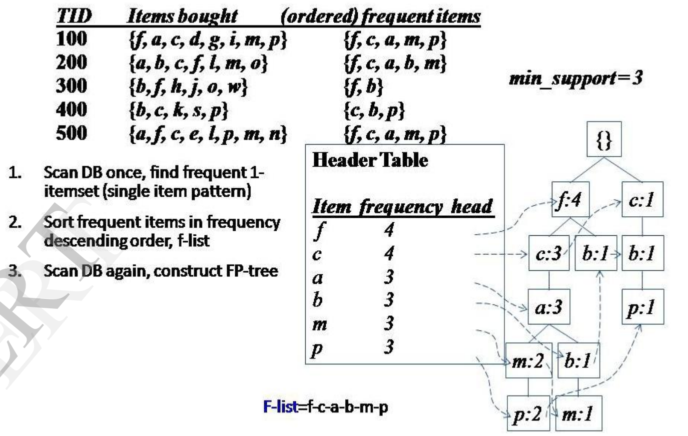
Also called Generalized Sequential Pattern, is an Apriori based data mining algorithm. It works as:

Pseudo Code:

For a given sample data, every item in DB is a candidate of length-1, in start:

```
for each level (i.e., sequences of length-k) do
        Go through database and collect support count for every candidate sequence
        Generate candidate length-(k+1) sequences from length-k frequent sequences using Apriori.
repeat as long as there are no frequent sequences or no candidate can be found.
```

Example:



TID	*Items bought*	*(ordered) frequent items*
100	*{f, a, c, d, g, i, m, p}*	*{f, c, a, m, p}*
200	*{a, b, c, f, l, m, o}*	*{f, c, a, b, m}*
300	*{b, f, h, j, o, w}*	*{f, b}*
400	*{b, c, k, s, p}*	*{c, b, p}*
500	*{a, f, c, e, l, p, m, n}*	*{f, c, a, m, p}*

min_support=3

1. Scan DB once, find frequent 1-itemset (single item pattern)
2. Sort frequent items in frequency descending order, f-list
3. Scan DB again, construct FP-tree

Header Table

Item	*frequency*	*head*
f	*4*	
c	*4*	
a	*3*	
b	*3*	
m	*3*	
p	*3*	

F-list=f-c-a-b-m-p

Example-1: Sequencing using GSP.[18]

4.5. Constraints on Sequencing:

Sequential data mining faces following constraints:

I. Item constraint (web log patterns only about e-stores)

II. Length constraint (patterns having at least 40 cloth items)

III. Super pattern constraint (super patterns of "Cannon digital camera")

IV. Aggregate constraint(patterns such that the avg. price of items is below $100)

5. CLASSIFICATION[10][14][16]

Classification based data mining exists as the backbone of machine learning in artificial intelligence. Classification generally consists of assigning a class label to a set of non-labeled cases [16]. The process of assigning a class or a label to unspecified data can be achieved by either of two ways: Supervised Classification and Unsupervised classification [10].

Supervised Classification: Given labeled data, predict output. Ie, set of possible classes is known/told via sample data.

Unsupervised Classification: Unlike supervised classification, sample data for unsupervised classification lacks predefined class labels. Hence its users responsibility to explore the data to find some intrinsicstructures (common features for classification) in them. Clustering is most widely used unsupervised classification technique. KNN and neural network are others.

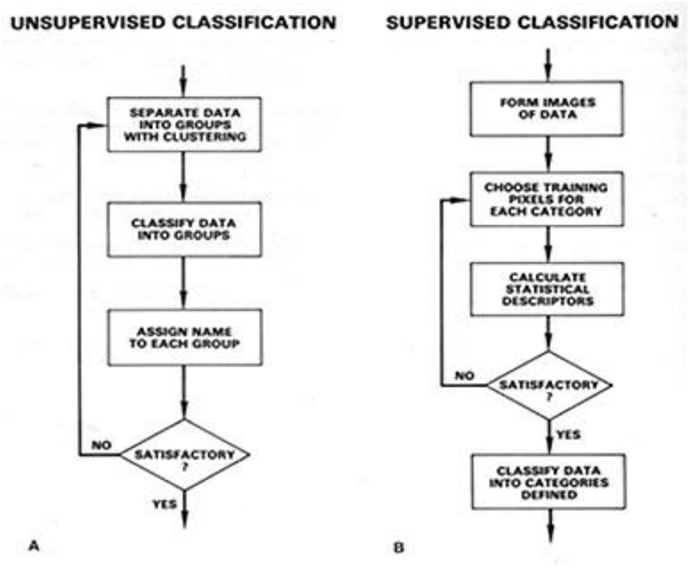

UNSUPERVISED CLASSIFICATION
SUPERVISED CLASSIFICATION
SEPARATE DATA INTO GROUPS WITH CLUSTERING
CLASSIFY DATA INTO GROUPS
ASSIGN NAME TO EACH GROUP
NO
SATISFACTORY ?
YES
A
FORM IMAGES OF DATA
CHOOSE TRAINING PIXELS FOR EACH CATEGORY
CALCULATE STATISTICAL DESCRIPTORS
NO
SATISFACTORY ?
YES
CLASSIFY DATA INTO CATEGORIES DEFINED
B

Diagram-1: Process Flow of Supervised vs. Unsupervised Classification [10]

5.1. Supervised Classification

5.1.1. Process:

I. Given sample data, also called training set, consists of multiple entries, each with multiple characteristics or features.
II. Each record is given a class label.
III. The purpose of classification is to analyze the sample data and to develop an accurate understanding or model for each class using the attributes present in the data.
IV. This model is then used to classify/label test data.

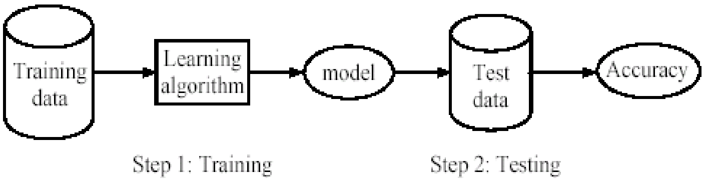


Diagram -2: Linear Overview of steps involved in Supervised Classification.[16]

5.1.2. Example:
We are given here an excerpt from university database as training data, with attributes including teacher's name, rank and years of experience with class label tenure with classes yes or no. Our purpose here is to analyze this data, learn patterns (using rows) and determine class for test data.
Test Data: *IF rank = 'professor' OR years > 6 THEN tenured? Seeing sample data, tenure be = 'yes'.*

Table-4: Sample Label Data of Professors.[16]

NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

5.2. Common Techniques of Supervised Classification:
Bayesian Classification, Naïve Bayesian Classification, Robust Bayesian Classifier, Decision tree learning etc.

5.3. Unsupervised Classification:
In an unsupervised classification, the objective is to group together sparse, visibly different response patterns into multiple clusters that are statistically separable. Thus, of given 50 data points, a small range of digital numbers (DNs) let's say 5 bands, can form one cluster that is different from another cluster of say 10 bands and so on. Separation or the distance between clusters will depend on the parameters (features) we choose to differentiate [10].

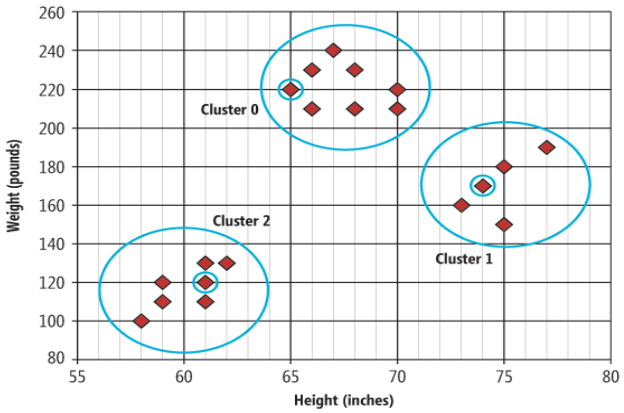
5.2. Common Techniques of Unsupervised Classification:
Clustering, K-Means Classification, KNN Classification, Decision tree learning etc.
Here we will discuss K-Means classification only.

5.3. K-Means Classification:
It is a Nonhierarchical unsupervised classification type where each instance is placed in exactly one of K non-overlapping clusters [16].Since only one set of clusters is output, hence it's unto user to input desired number of clusters K beforehand.

5.3.1. Algorithm K-Means:

I. Decide the value for K.
II. Initialize the *k* cluster centers (randomly, if necessary), each consisting of multiple entries, with multiple characteristics or features.
III. With given record/data points. Decide class memberships of the *N* objects by assigning them to the nearest cluster center.
IV. Re-calculate k cluster centers, assuming that the memberships determined above are correct.
V. If none of the *N* objects changed membership in the last iteration, exit. Otherwise goto III.

Example: In the graph below, raw data of students of Grade IX were given, based on K-3 (where diamond in little blue circle represents k value), three clusters are formed.



Graph -3: Clustering using K-3[14].

More on K-Means:

I. Relatively efficient: O(tkn), where n is no of objects, k is no of clusters and t is no of iterations. Normally k.t<<n.

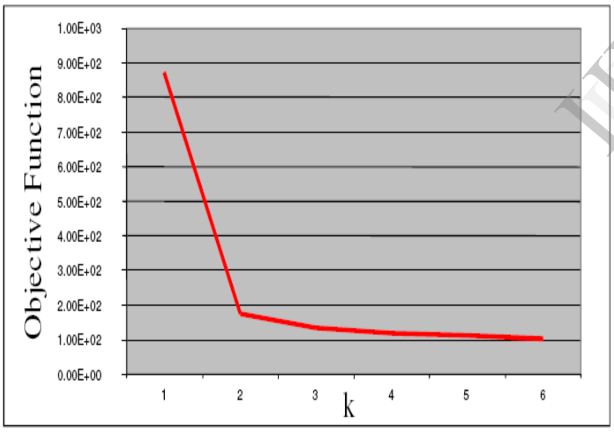
II. Terminates often at local optima.

III. Applicable only when a mean is defined, i.e. not very effective for categorical data.

IV. K also should be defined in start, i.e. no of clusters.

V. Finds faults with noisy data and outliers, hence other means should be used to minimize noise before using k-means.

VI. Value of K directly affect Objective function, hence optimum value of K must be used.



Graph -4: Objective function on different values of K[17].

Where, objective function is directly related to squared error that is prone to occur. (2^{nd} formula represents objective function)

$$se_{K_i} = \sum_{j=1}^{m} ||t_{ij} - C_k||^2$$

$$se_K = \sum_{j=1}^{k} se_{K_j}$$

m = Number of data-points in one cluster
K = total number of clusters

6. CLUSTERING[9][17]

Clustering, an example of unsupervised learning is defined as the ‘grouping of similar data points’ from raw, unlabeled data based on some common features. This is often achieved by comparing similarity of items in sample data. Clustering at times is considered as largely subjective.

Hard Vs. Soft Clustering:
Clustering is considered as hard if a single data item can belong to any one cluster, however in soft clustering a single data point is allowed to be a part of two clusters simultaneously [9].

Examples:
Hard Clustering: K-Means clustering, K-Means ++
Soft Clustering: Gaussian mixture model

6.1. Types of Clustering:

Partitional Algorithms: Construct various partitions and then evaluate them by some criterion (K-Means, K-Means++)

Hierarchical Algorithms: Make a hierarchical decomposition of the set of objects using some criterion (Agglomerative, single link).

Here we will discuss both Agglomerative also called Bottom-Up clustering and Divisive also called Top Down Clustering from Hierarchical clustering:

6.2. Agglomerative Clustering:

Starting with each item in its own cluster, purpose is to find the best pair to merge into a new cluster at each respective level. While repeating until all clusters are fused together.

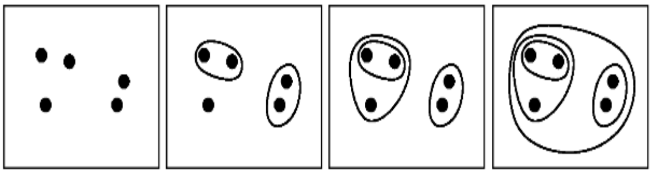


Diagram -3: Step Wise merging of data points using Agglomerative Clustering[17].

6.3. Divisive Clustering (Top Down Clustering):

Starting with all data points in one cluster, the cluster in each step then splits up using a flat clustering algorithm. The process is applied recursively until all data points exist in their own singleton cluster.

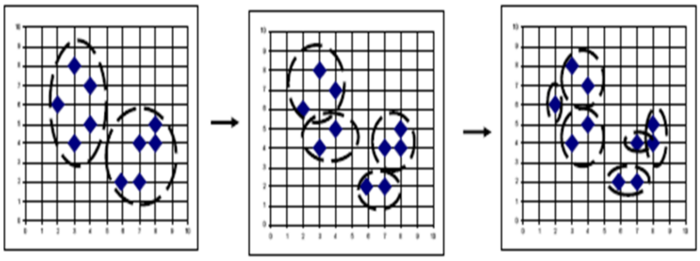


Diagram -4: Step Wise splitting of data points using Bottom up Clustering[17].

6.4. Gaussian Mixture Model (Soft Clustering)[9]:

Gaussian mixture model finds its large scale usage in image processing**.** The Gaussian mixture architecture measures probability density functions (PDF) for each given class, and then performs relative classification based on Bayes' rule:

$$P(C_i \mid X) = P(X \mid C_i) \cdot \frac{P(C_j)}{P(X)}$$

Where P(X | C_i) is the PDF of class j, evaluated at X, P(C_j) is the prior probability for class j, and P(X) is the overall PDF, evaluated at X.

Unlike the normal or uni-modal Gaussian architecture, which assumes P(X | C_j) to be in the form of a Gaussian, the Gaussian mixture model uses P(X | C_j) as a weighted average of multiple Gaussians.

$$\text{i.e.}\quad P(X \mid C_j) = \sum_{k=1}^{Nc} w_k G_k$$

Where w_k is the weight of the k-th Gaussian G_k, and all weights accumulate to sum one. Hence, using this one such PDF model is produced for each class. Where each Gaussian component is defined as:

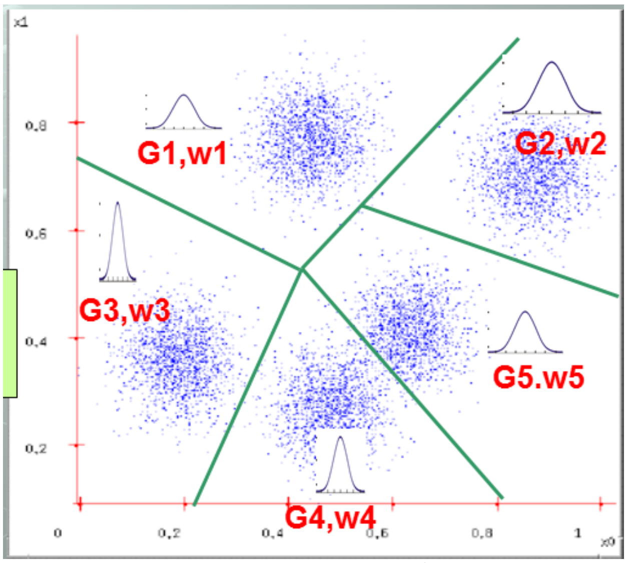
$$G_k = \frac{1}{(2\pi)^{n/2} \left| V_k \right|^{1/2}} \cdot e^{[-1/2(X-M_k)^T V_k^{-1}(X-M_k)]}$$

Where M_k is the mean of the Gaussian and V_k is the covariance matrix of the Gaussian.

Whereas the free parameters of the Gaussian mixture model consist of the means and covariance matrices of the Gaussian components and the weights indicating the contribution of each Gaussian to the approximation of P(X | C_j).

Hence using above three formulas with variables: μ_i**,** V_i**,** w_k

We use EM (estimate-maximize) algorithm to approximate this variables.



Graph -5: Gaussian mixture model of Class 1[9].

Defined parameters are tuned using a complex iterative procedure called the estimate-maximize (EM). It's simply an algorithm, which aims at maximizing the likelihood of the training set generated by the estimated PDF.

L, the likelihood function for each class j can be defined as:

$$L_j = \prod_{i=0}^{N_{train}} P(X_i \mid C_j) \Longrightarrow \sum_{i=0}^{N_{train}} \ln(P(X_i \mid C_j))$$

6.4.1. Gaussian Mixture Training Flow Chart

I. Using K means clustering algorithm, initialize the initial Gaussian means μ_i, i=1,…G**.**

II. Initialize $V_{i,,,}$ the covariance matrices, to the distance to the nearest cluster.

III. Initialize the weights π_i=1 / G so that all Gaussian are equally likely

IV. Using formula, present each pattern X of the training set and model each of the classes K as a weighted sum of Gaussians.

$$p(X \mid \theta_s) = \sum_{i=1}^{G} \pi_i \, p(X \mid G_i)$$

Where **G**is the number of Gaussians, and $\boldsymbol{\pi_i}$**'s** are the new weights, whereas $\mathbf{V_i}$is the covariance matrix.

$$p(X \mid G_i) = \frac{1}{(2\pi)^{d/2} \left| V_i \right|^{1/2}} \cdot e^{[-1/2(X-\mu_i)^T Vi^{-1}(X-\mu_i)]}$$

V. Compute:

$$\tau_{ip} \equiv P(G_i \mid X) = \frac{\pi_i \, p(X \mid G_i, C_k)}{p(X)} = \frac{\pi_i \, p(X \mid G_i, C_k)}{\sum_{j=1}^{G} \pi_j \, p(X \mid \theta_j, C_k)}$$

VI. Iteratively update the weights, means and covariance using:

$$\pi_i(t+1) = \frac{1}{N_c} \sum_{p=1}^{N_c} \tau_{ip}(t) \text{ for weights.}$$

$$\mu i(t+1)=\frac{1}{N_c\pi_i(t)}\sum_{p=1}^{N_c}\tau_{ip}(t)X_p$$ for means. And

$$V_i(t+1)=\frac{1}{N_c\pi_i(t)}\sum_{p=1}^{N_c}\tau_{ip}(t)((X_p-\mu_i(t))(X_p-\mu_i(t))^T)$$ for variance.

VII. Recalculateτ_{ip}using the new weights, new means and covariance. And quit training if

$$\Delta\tau_{ip}\equiv\tau_{ip}(t+1)-\tau_{ip}(t)\le threshold$$

Or the number of epochs reaches the specified value. Otherwise, continue the iterative process.

VIII. Finally, present each input pattern X and compute the confidence for each class j:

$$P(C_j)P(X\,|\,\theta_x,C_j)$$

Where $P(C_j)=\frac{N_{ci}}{N}$ is the prior probability of class **C**$_{j,}$ being estimated by counting the number of training patterns and classify pattern X as the class with the highest confidence.

6.5. Summary:

I. Time complexity: O(n^2), where *n* is the number of total objects.

II. Like any heuristic search algorithms, local optima are a problem here.

III. Interpretation of results is (very) subjective.

7. KNN[14][17]

KNN also called K-Nearest neighbors is a popular uninformed data mining technique. Given training data (X(1),D(1)), (X(2),D(2)), …, (X(N),D(N)), We first define a distance metric to measure distance between points in inputs space.

Commonly used distance measure is '**Euclidean Distance'**:

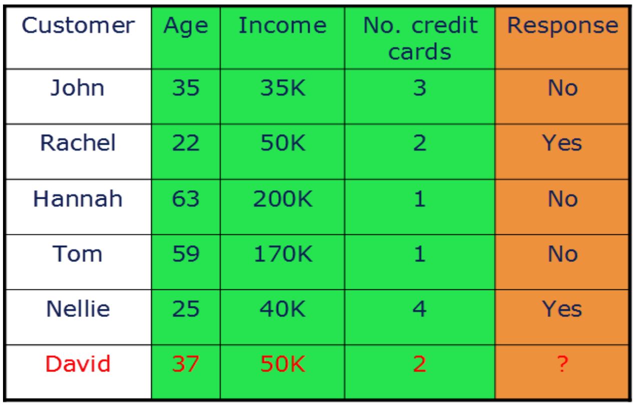
$$D(i,j)=\sum_{k=1}^{n}\left(x_k(i)-x_k(j)\right)^2$$

7.1. Working:

Given a test point X. Our purpose is to find the K nearest training inputs to X. Mark these points as: (X(1),D(1)), (X(2), D(2)), …, (X(k), D(k)). Classification then of point X is carried out via: Y = most common class in set {D(1), D(2), …, D(k)} and X->D.

7.2. Example:

Classify whether a customer will respond to a survey question using a 3-Nearest Neighbor classifier.



Customer	Age	Income	No. credit cards	Response
John	35	35K	3	No
Rachel	22	50K	2	Yes
Hannah	63	200K	1	No
Tom	59	170K	1	No
Nellie	25	40K	4	Yes
David	37	50K	2	?

Table-5: Sample Customer Data[17].

Using Euclidian distance, calculating all customers distance from David gives us John, Rachel and Neile as 3 nearest neighbours, at distances of (15.16,No), (15,Yes),(15.74,Yes) respectively.Hence Response value for David be Yes, based on majority.

7.3. Pseudo Code for KNN:

Training Algorithm:

I. Store all training examples *<x, f(x)>*

II. Find best value for K

Classification Algorithm:

Given a query instance x_q to be classified,

Let x_1, …x_k denote the k instances from the list of training examples

Return $$\hat{f}(x_q)\leftarrow \arg\max\sum_{i=1}^{k}\delta(c,f(x_i))$$

where δ(a,b)=1 if a=b and where δ(a,b)=0 otherwise

(C = class)

7.4. Complexity:

O(Nd) for both storage and query time where N is the number of training examples, and d is the dimension of each sample.

7.5. Variations of KNN:

K-Means, K-Means++,Matching Matrix are few widely used variations of KNN.

8. CONCLUSIONS

Today data mining is widely used in diverse areas. There is hardly any walk of life where data mining finds hard to find its application and usage. A number of commercial data mining systems are available today yet there are many challenges in this field. Broadly, we can classify data mining applications based on their usage into: Financial Data Analysis, Retail Industry, Telecommunication Industry, Biological Data Analysis, Other Scientific Applications and Intrusion Detection [15].

However the choice for which data technique to be used depends largely on your type of data, dimensions, data existing state/stage, and the in-house data state.

From the discussion above its obvious that association based techniques provides patterns of associated values of variables and frequencies of their appearance. Similarly, classification provides means to use data for prediction and future responses. Clustering whereas provides grouping of homogeneous objects. Based on certain hypothesis about number of classes to be found; as in KNN. Results are directly understandable [1].However they normally do not work well with very big data sets.

Another important thing is the algorithm one implies in particular mining technique, since it's the working engine that's defines running of a car.

REFERENCES

[1]. K.Gibert, M.Sanchez&V.Codian. Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation, http://www.iemss.org/iemss2010/papers/S23/S.23.03.Choosing%20the%20Right%20Data%20Mining%20Technique%20%20Classification%20of%20Methods%20and%20Intelligent%20Recommendation%20-%20MIQUEL%20SANCHEZ-MARRE.pdf

[2]. P.Adriaans and D.Zantinge. Data Mining. Addison Wesley Longman,Harlow, England, 1996.

[3]. J.Han and M.Kamber. Data Mining: Concepts and Techniques.Morgan Kaufmann, San Francisco, CA, 2000.

[4]. C. Schoenbach, P. K.Saunders, and V. Brusic. Data warehousing inmolecular biology. Briefings in Bioinformatics, 1:190–198, 2000.

[5]. O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA Microarrays. Bioinformatics, 17:520–525, 2001.

[6]. R. Sandy. Statistics for Business and Economics. McGrawHill,1989.

[7]. I. T. Jolliffe. Principal Component Analysis. Springer Verlag, Berlin, 1986.

[8]. U. Fayyad and K. Irani. Multi-interval discretization of continuous-valued at-tributes for classification learning. In Proceedings of 13th International Joint Conference on Artificial Intelligence, pages 1022–1029, 1993.

[9]. K.Hsien. (June 2005). K Means Clustering, Nearest Cluster and Gaussian Mixture, http://www.ibms.sinica.edu.tw/~pan/classification/documents/Gaussian%20Mixture,%20Nearest%20Cluster%20and%20K%20means.ppt

[10] Classification, http://ces.iisc.ernet.in/hpg/envis/Remote/section27.htm

[11]. N.Cerpa.Support vs. Confidence in Association Rule Algorithm.http://www.academia.edu/648890/Support_vs_Confidence_in_Association_Rule_Algorithms

[12]. CS583, Bing Liu, UIC. Mining Association Rules. http://www.cs.uic.edu/~liub/teach/cs583-fall-05/CS583-association-rules.ppt

[13]. J.Pei, H.Pinto&J.Han.Prefix span: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth.http://dm.kaist.ac.kr/kse525/resources/papers/icde2001prefixspan.pdf

[14]. J.McCaffrey. DetectingAbnormal DataUsing k-Means Clustering.http://msdn.microsoft.com/en-us/magazine/jj891054.aspx.

[15] Data Mining - Applications & Trends.http://www.tutorialspoint.com/data_mining/dm_applications_trends.htm.

[16]. A brief summary of classification techniques. http://www.cs.zju.edu.cn/people/xucf/course/DM2007/Classification.ppt

[17].Unsupervised Learning, Data mining and knowledge discovery.http://www.public.iastate.edu/~olafsson/unsupervised_learning.ppt

[18] Data Mining. http://www.rainasolutions.org/p/data-mining_7438.html

[19].http://ces.iisc.ernet.in/hpg/envis/Remote/section27_files/Unsup-SupClassif.jpeg