DISCOVERY IS THE PROCESS OF ANALYSIS OF DATA FROM DIFFERENT VIEWS AND PERSPECTIVES

Dr.N K Sirohi[1], Gauri Sharma [2]

[1,2]*Dept. of Electronics Engineering, SESCOE, Navalnagar (India)*

ABSTRACT

We are living in data age where amount of data are collected daily.analyzing such data is a important need. Nowadays there is huge amount of data being collected and stored in databases everywhere across the globe. Throughout the years many algorithms were created to extract of knowledge from large sets of data. Density estimation is the ubiquitous base modelling mechanism employed for many tasks including clustering, classification, anomaly detection and information retrieval. Commonly used density estimation methods such as kernel density estimator and k-nearest neighbour density estimator have high time and space complexities The Bayesian algorithm totally dependent on density estimation for the base modelling so all the algorithms those depends on Bayesian algorithm is not suitable for big or large data that's why we use the Mass based classifiers, which is based on the mass estimation. Mass estimation a base modelling mechanism that can be employed to solve various tasks in machine learning. mass estimation solves problems effectively in tasks such as information retrieval, regression and anomaly detection. The models, which use mass in these three tasks, perform at least as well as and often better than eight state-of-the-art methods in terms of task-specific performance measures. mass estimation has constant time and space complexities .Mass estimation possesses different properties from those offered by density estimation.

***Keywords**: Density Estimation, Kernel Density Estimator-Nearest Neighbour Density Estimator, Mass Estimation, Mass Distribution, Binary Splits.*

I INTRODUCTION

Data mining sometimes called data or knowledge discovery is the process of analysis of data from different views and perspectives and then summarizing them into useful information .The information can be used to increase revenue, cuts costs, or both. Data mining software consists number of analytical tools for analyzing data.Data classification is the categorization of data for its most effective and efficient use for many applications. In a basic approach to storing computer data, data can be classified according to its critical value or how often it needs to be accessed, with the most critical or often-used data stored on the fastest media while other data can be stored on slower (and less expensive) media. Data mass or mass is defined as a no of points or data instances in a region or closed region. Application of mass is used to solve the various data mining problems such as information retrieval, clustering, regression, anamoly detection. Data mining methods which are based on the mass sometimes performed better than other state of arts methods.mass is not a probability function,it does not provide probability.Mass in a given region or space defined using a rectangular function.The mass for the point is

calculated from average of masses from all regions.generate possible regions using binary spilt and generate random axis parellal regions.mass estimation is more effective due to mass compution require simple counting.Mass distribution viewed as ordering from core points to the fringe points in a data cloud.mass distribution estimation uses binary splits in a one dimensional region which seprate the space into two non empty Regions.it uses half space trees for a multidimensional mass estimation.

II RELATED WORK

2.1 Density Estimation

This describes probably three most commonly used density estimation methods, namely KDE, k -nearest neighbour density estimator and neighbourhood density estimator.

2.1.1 Kernel density estimator

Let **x** be an instance in a *d*-dimensional space R_d . The KDE [1], defined by a kernel function $K(\cdot)$ and bandwidth *b* is given
as follows ;

$$\bar{f}_{\mathrm{KDE}}(\mathbf{x}) = \frac{1}{nb^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{b}\right) \quad (1)$$

The difference between $\mathbf{x} - \mathbf{x}_i$ requires some form of distance measure which calculate the distance.Here *n* is the number of instances in the given data set *D*. An example of $K(\cdot)$, as a rectangular function, is given as follows;

$$K(\mathbf{x}) = \begin{cases} \frac{1}{2} & \text{if } |\mathbf{x}| < 1 \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

2.1.2 k-NN density estimator

A *k*-NN density estimator [2] can be expressed as follows ;

$$\bar{f}_{\mathrm{kNN}}(\mathbf{x}) = \frac{|N(\mathbf{x}, k)|}{n \sum_{\mathbf{x}' \in N(\mathbf{x},k)} \mathrm{distance}(\mathbf{x}, \mathbf{x}')} \quad (3)$$

Where $N(\mathbf{x}, k)$ is the set of *k* nearest neighbours to **x**, and the search for nearest neighbours [3],[4] is conducted over *D* of size *n*.

2.2 Mass Estimation

MassBayes based on the mass estimation or mass distribution. The mass for the point is calculated from average of masses from all regions.generate possible regions using binary spilt and generate random axis parellal regions.mass estimation is more effective due to mass compution require simple counting. Mass distribution viewed as ordering from core points to the fringe points in a data cloud.mass distribution estimation uses binary splits in a one dimensional region which seprate the space into two non empty Regions.it uses half space trees for

a multidimensional mass estimation.Density estimation [5], is expensive in terms of time and space complexities than mass estimation.

2.3 Mass Distribution Estimation Using Binary Splits

Divide the data set into two seprate regions and calculate the mass in each region.the mass distribution at point x is estimated to be the sum of all the masses (weights) from regions which are occupied by x, as a result n-1 binary splits for a data set of size n.Let $x_1 < x_2 < \cdots < x_{n-1} < x_n$ on the real line, $x_i \in R$ and $n > 1$. Let s_i be the binary split between x_i and x_{i+1}, yielding two non-empty regions having two masses m_i^L and m_i^R.

Mass base function: $m_i(x)$ as a result of s_i,is defined as;

$$m_i(x) = \begin{cases} m_i^L & \text{if } x \text{ is on the left of } s_i \\ m_i^R & \text{if } x \text{ is on the right of } s_i \end{cases} \quad (4)$$

Mass distribution: mass(x_a) for a point $x_a \in \{x_1, x_2, \ldots x_{n-1}, x_n\}$ is defined as summation of a series of mass base functions $m_i(x)$ weighted by $p(s_i)$ over $n-1$ splits as follows, where $p(s_i)$ is the probability of selecting s_i.

$$\begin{aligned} mass(x_a) &= \sum_{i=1}^{n-1} m_i(x_a)\, p(s_i) \\ &= \sum_{i=a}^{n-1} m_i^L\, p(s_i) + \sum_{j=1}^{a-1} m_j^R\, p(s_j) \\ &= \sum_{i=a}^{n-1} i\, p(s_i) + \sum_{j=1}^{a-1} (n-j)\, p(s_j) \end{aligned} \quad (5)$$

2.4 Level-h Mass Distribution Estimation

Level-h mass distribution estimation viewed as a localised versions of the basic level-1 estimation.The level-h mass distribution for a point $x_a \in \{x_1, \ldots, x_n\}$, where $h < n$, is expressed as: Here a high level mass distribution is computed recursively by using the mass distributions obtained at lower levels. A binary split s_i in a level-h(> 1) mass distribution produces two level-(h-1) mass distributions: (a) *mass* L_i $(x,h\text{-}1)$—the mass distribution on the left of split s_i, which is defined using $\{x_1, \ldots, x_i\}$; and (b) *mass*R_i $(x,h\text{-}1)$.the mass distribution on the right which is defined using $\{x_{i+1}, \ldots, x_n\}$. Equation is the mass distribution at level-1.

2.5 Multi Dimensional Mass Estimation

Multi dimensional mass estimation , diminish the need to probability computation of a binary split.generate the multiple regions which are randomly selected,which cover a point, then mass is calculated by averaging all masses from all regions.the random regions are generated using axis-parellal splits called half space splits.each half space splits performed on randomly generated attributes.

For a h-level split, each half-space split is carried out h times recursively along every path in a tree structure. Each h-level (axis-parallel) split generates $2h$ non-overlapping regions or spaces. Multiple h-level splits are used to estimate mass for each point in the feature space.The multi-dimensional mass estimation requires two functions. First, it require a function that generates random regions that covering each point in the feature space. This

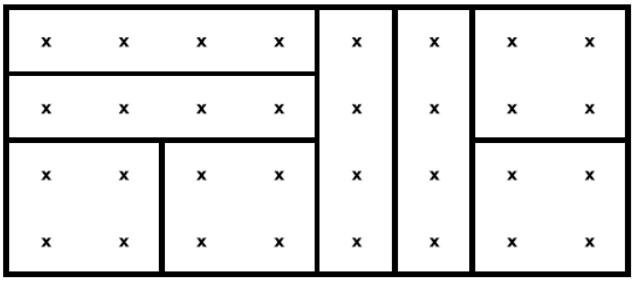
function is a generalisation of the binary split into half-space splits or 2*h*-region splits when *h* levels of half-space splits are used. Second, a generalised version of the mass base function is used to define mass in a region. Let **x** be an instance in *Rd*. Let *T h(***x***)* be one of the 2*h* regions in which **x** falls into; *T h(·)* is generated from the given data set *D*, and *T h(·|D)* is generated from $D \subset D$; and m be the number of training instances in the region. The generalised mass base function: m(T h(x)) is defined as;

T(h) is defined in multi-dimensional space, the multi-dimensional mass estimation

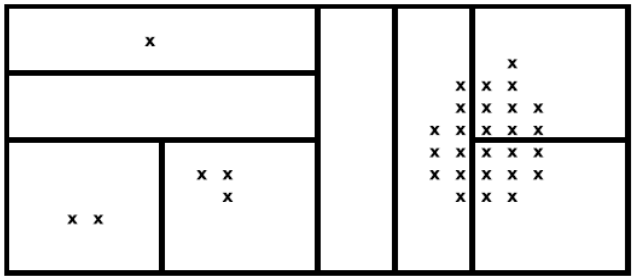
$$\mathbf{m}(T^h(\mathbf{x})) = \begin{cases} \mathsf{m} & \text{if } \mathbf{x} \text{ is in a region of } T^h \text{ having } \mathsf{m} \text{ instances} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

2.6 Half-Space Trees for Mass Estimation

We use the half space tree to implement the multi dimensional mass estimation. Half-Space Tree, comes from the fact that equal size regions contain the same mass in a space uniform mass distribution, regardless of the shapes of the regions. This is shown in Fig. 2(a), where the space enveloped by the data is split into equal-size half-space recursively three times into eight regions. Note that the shapes of the regions may be differ from each other because the splits at the same level may not use the same attribute.The binary half-space split ensures that every split produces two equal-size half-spaces,each containing exactly half of the mass before the split under a uniform mass distribution.This characteristic enables us to compute the relationship between any two regions easily.For example, the mass in every region shown in Fig. 2(a) is the same, and it is equivalent to the original mass divided by 23 because three levels of binary half-space splits are applied. A deviation from the uniform mass distribution allows us to rank the regions based on mass. Figure (b) provides such an example in which a ranking of regions based on mass provides an order of the degrees of anomaly in each region.



(a) Uniform mass distribution.



(b) Non-uniform mass distribution.

Fig. 2. Half space subdivisions of (a) uniform distribution (b) non uniform distribution

HS-Tree: based on mass only. The first variant, HS-Tree, builds a balanced binary tree structure which makes a half-space split at each internal node and all external nodes have the same depth. The number of training instances falling into each external node is recorded and it is used for mass estimation.

HS*-Tree: based on augmented mass. Unlike HS-Tree, the second variant, HS*-Tree, whose external nodes have differing depth levels.

In a special case of HS*-Tree, the tree growing process at a branch will only terminate to form an external node if the training data size at the branch is 1 (i.e., the size limit is set to 1).

Under uniform mass distribution, the mass at level i is related to mass at level 0 as follows;

$$m[0] = m[i] \times 2^i , \quad (7)$$

or mass values between any two regions at levels i and j related as follows;

$$m[i] \times 2^i = m[j] \times 2^j \quad (8)$$

Under non-uniform mass distribution, the following inequality establishes an ordering between any two regions which are lie in different levels:

$$m[i] \times 2^i < m[j] \times 2^j \quad (9)$$

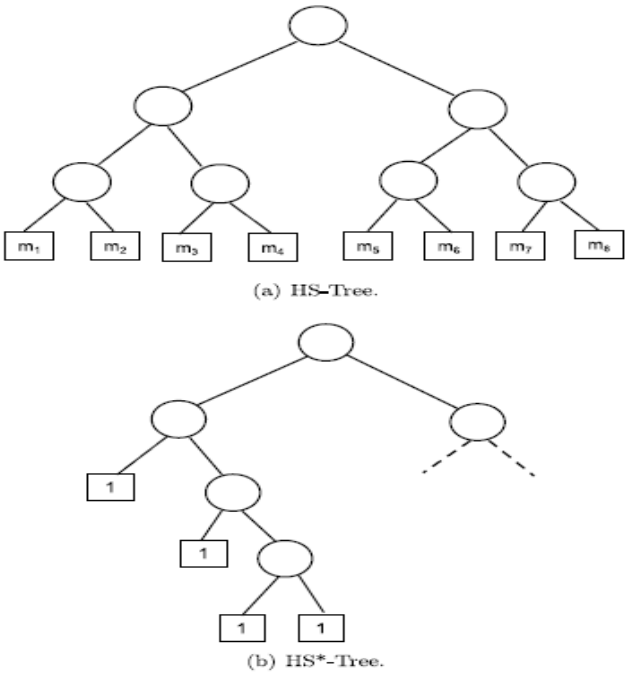


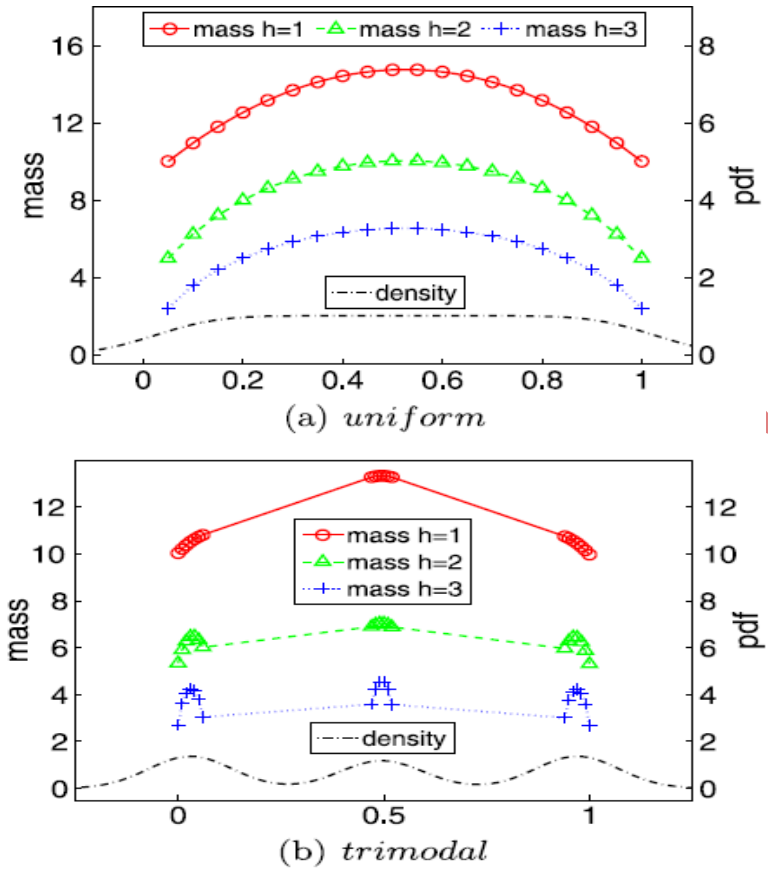
Fig. 3 Half-Space Tree: (a) HS-Tree: An HS-Tree for the data shown in Fig. 2(a) .
(b) HS*-Tree: An example of a special case of HS*-Tree when the size limit is set to 1.

III SIMULATION RESULTS

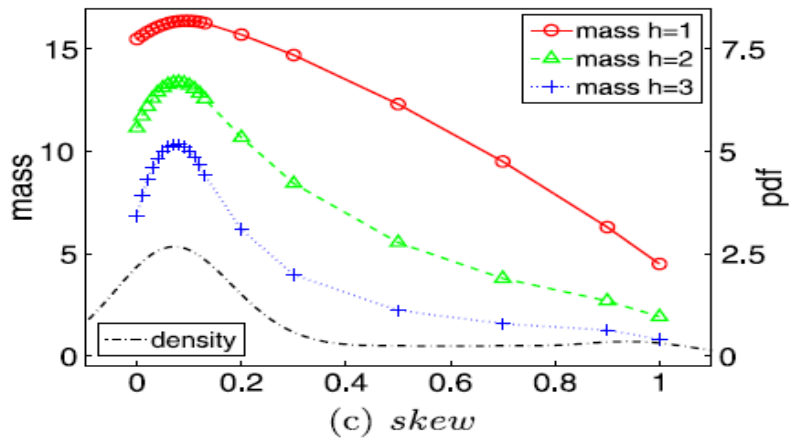
We adopt the simulation software mat lab and set up the scenario for experiment.Density estimation is the ubiquitous base modelling mechanism employed for many tasks including clustering, classification, anomaly detection and information retrieval.Commonly used density estimation methods such as kernel density estimator and k-nearest neighbour density estimator have high time and space complexities which render them inapplicable in problems with big data. This weakness sets the fundamental limit in existing algorithms for all these tasks.

3.1 Level-h Mass Distribution

Estimate the mass distribution for h=1,2,3 and density distribution from kernel density estimation with 0.1 bandwidth .The dataset have 20 points.



The h=1 mass estimation looks at the data as a group and produce a concave function.the fig (c) has a skew density distribution,the distinction between near fring points and far fringe points.



Mass estimation possesses different properties from those offered by density estimation.A mass distribution stipulates an ordering from core points to fringe points in a data cloud. This ordering accentuates the fringe points with a concave function derived from data, resulting in fringe points having markedly smaller mass than points

close to the core points. Mass is computed by simple counting and it requires only a small sample through an ensemble approach.

IV CONCLUSION

Mass estimation has two important advantages. First, the concavity property mentioned above ensures that fringe points are 'stretched' to be farther from the core points in a mass space making it easier to separate fringe points from those points close to core points. This property in mass space can then be exploited mass estimation offers to solve a ranking problem more efficiently using the ordering derived from data directly without expensive distance (or related) calculations. The analysis and comparison of the density estimation and Mass estimation in one dimensional space and multi dimensional space in terms of their performance and other parameters.the density estimation and its other classification algorithms is not suitable for large data set in terms of time and space complexities so we use the mass estimation. Mass estimation has two advantages in relation to efficacy and efficiency. First, the concavity property mentioned above ensures that fringe points are 'stretched' to be farther from the core points in a mass space making it easier to separate fringe points from those points close to core points. This property in mass space can then be exploited mass estimation offers to solve a ranking problem more efficiently using the ordering derived from data directly without expensive distance (or related) calculations.

REFERENCES

[1.] Goldstein J, Ramakrishnan R, Shaft U (1999) When is "nearest neighbor" meaningful? In: Proceedings of the 7th international conference on database theory, pp 217–235

[2.] Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapmal & Hall/CRC (1986)

[3.] Beygelzimer A, Kakade S, Langford J (2006) Cover trees for nearest neighbor. In: Proceedings of the 23rd international conference on machine learning, pp 97–104

[4.] Kriegel H-P, Ng RT, Sander J (2000) LOF: identifying density-based local outliers.In: Proceedings of ACM SIGMOD international conference on management of data, pp 93–104

[5.] Ting, K. M., & Wells, J. R. (2010). Multi-dimensional mass estimation and mass-based clustering. In *Proceedings of IEEE ICDM* (pp. 511–520).