Internet of Things: Comparative Study on Classification Algorithms (k-NN, Naive Bayes and Case based Reasoning)

Roshna Chettri
Sikkim Manipal Institute of Technology.
Sikkim Manipal University

Shrijana Pradhan
Sikkim Manipal Institute of Technology.
Sikkim Manipal University

Lekhika Chettri
Sikkim University
Gangtok

ABSTRACT
The Internet of Things aims to connect each existing things in the world with internet. It is sweeping all the things to a world like a garland where each flower is connected by a sting forming connectivity. IoT can be considered as a big bucket where everything's, every data in the world can be poured to form a live-like connectivity, and hence needs data computation for prediction of the unknown data. Data computation in the internet of things is incorporated to return the value data from the huge collected data collected from the different sources device of the IoT. There are various algorithms for computation of data. This paper focus on comparing supervised learning algorithms i.e. K-NN, Naive Bayes and Cased Based Reasoning (CBR) Classifier.The effects of the mentioned algorithms are based on the following parameters i.e. size of the dataset, performance, processing time and accuracy.

Keywords
K-NN, Naive Bayes, Case Based Reasoning.

1. INTRODUCTION
In classification, a classifier is learned from a set of training examples with class labels. The performance of a classifier is determined by its classification accuracy. Classification techniques in data mining are capable of processing a large amount of data. It can predict class labels and classifies data based on training set and class labels and therefore can be used for classifying unseen available data.

2. CLASSIFICATION ALGORITHM

2.1 K- Nearest Neighbors Classification
K Nearest Neighbors is a simple algorithm that stocks all existing cases and classifies new cases/data based on a similar measure (e.g., distance functions). It has been widely used for decades as an effective classification mode. It assumes that the data is in a feature space. The data can be scalars or possibly even multidimensional vectors. K decides how many neighbors (neighbor is defining on the base of distance metric) influence the classification. [5] This is usually an odd number say if the number of classes is 2 then k=1 and the algorithm is simply called the nearest neighbor algorithm. Choice of k is very acute [8] – A small value of k means that noise will have a higher influence on the result. A large value makes it computationally expensive and can defeats the basic philosophy behind KNN.

Description of an algorithm
Example1: In this example a factory produce a new paper so without survey we can identify whether that paper will be good or bad. Using k-NN the following steps are followed to identify paper classification.

2.1.1 Training set: (x1, y1), (x2, y2), …, (xn,yn)
The training data are stored in advance. This data is used for determining the accuracy of the algorithm and classifying. Referring to this training data unknown data are identified and are classified.

Table 1: Data set for training data

X1= Acid durability	X2= Strength	Y=Classification
5	5	Bad
8	5	Bad
3	5	Good
7	5	Good

Consider x1=3 and x2= 7.

K=3.

2.1.2 Calculate distance:
Calculate the distance between the query-instances and all the training data.

Table 2: Distance Calculation.

Distance calculation $(X-x1)^2+(x2- X2)^2$
$(5-3)^2 +(5-7)^2 = 8$
29
4
20

2.1.3 Determine nearest neighbor
Sort the distance to determine the nearest neighbor based on the kth minimum distance.

Table 3: Sorting Nearest Neighbor

Distance calculation	Rank minimum distance	Is it included in 3 nearest neighbor
8	2	Yes
29	4	No
4	1	Yes
20	3	Yes

2.1.4 Gather the category Y of the nearest neighbor

Table4: category Y of the nearest neighbor.

X= Acid durability	X2= Strength	Distance calculation	Rank minimum distance	Is it included in 3 nearest neighbor	Y
5	5	8	2	Yes	Bad
8	5	29	4	No	-
3	5	4	1	Yes	Good
7	5	20	3	Yes	Good

2.1.5 Use simple majority of the category of the nearest neighbors as the prediction of the query instance

We have seen that there are 2 good and 1 bad.

Predications is based on majority of votes and since there are two votes for good and one vote for bad. Therefore it concludes that paper the factory produce is good.

2.2 Naive Bayes classifier

Naive Bayes classification aim is to construct a rule which will allow assigning future objects to a class, given only the vectors of variables describing the future objects. To use this to produce classifications, we need to early estimation and training dataset. After that calculate all the possible probability of the given dataset. When new data arrive for classification then with the reference to priori calculated probability and training dataset it provide the final result indicating which class does the unseen data fall.

Description of algorithm

Example2: In this example for given dataset it tries to find either a tuple x buys a mobile phone or not.

2.2.1 D: Set of tuples

Each Tuple is an 'n' dimensional attribute vector

X: (x1, x2, x3,… ,xn)

X = (income = medium, credit rating = fair)

Table1: Dataset.

Record no	Credit rating	Income	Buys mobile
R1	Good	High	Yes
R2	Fair	Medium	No
R3	Excellent	High	Yes
R4	Fair	Medium	Yes

2.2.2 Determined a priori probability for each class.

Probability for a feature vector is calculated.
X = (income = medium, credit rating = fair)

P(X/Ci) = P(x1/Ci) * P(x2/Ci) *…* P(xn/ Ci).

P (c1) = (P (C1) = P (buys mobile = yes) =3/4=0.75
P (C2) = P (buys mobile = no) =1/4=0.25
P (income=medium / buys mobile = yes) = 1/3=0.33
P (income=medium / buys mobile = no) = 1/1=1
P (credit rating=fair / buys mobile = yes) = 1/3=0.33
P (credit rating=fair / buys mobile = no) = 1/1=1

P(X/ buys mobile = yes) = P (income=medium/ buys mobile = yes) * P (credit rating=fair/buys mobile = yes)
= 0.33
P(X/ buys mobile = No) =0.33

*2.1.3 Find class Ci that Maximizes P(X/Ci) * P(Ci)*

P(X/ buys mobile = yes) * P (buys mobile = yes) = 0.24
P(X buys mobile = No) * P (buys mobile = no) = 0.08

2.2.4: Prediction

Since probability of buying mobile is 0.24 which is greater than of not buying i.e. 0.08. Therefore a person belongs to tuple x will buy mobile.

2.3 Case Based Reasoning

In cased based reasoning it classifies the case according to the past experiences. [4]There are two ways of case-based reasoning: problem solving and interpretive. Problem solving method use old solution as the solution to the new problem and it provide warning message as well.[4]In the interpretive new situation are examine in the context of old situations After the problem situation has been assessed, the best matching case is searched in the case base and an approximate solution is retrieved. Retrieved solution is considered to fit for the better new problem. The adapted solution can be evaluated either before the solution is applied to the problem or after the solution has been applied. In any case, if the accomplished result is not reasonable, the retrieved solution must be adapted again or more cases should be retrieved. If the solution was verified as correct, the new case may be added to the case base.

It can be seen as the cycle of the following four tasks:

RETRIEVE:

In this phase similar case from the case based is retrieved for a given new case.

REUSE:

To fit into the new case adapt the retrieved case.

REVISE:

Estimate the result and revise it based on how well it works.

RETAIN:

Decision is made based either to retain the new case in the database or not.

3. COMPARING K-NN, NAIVE BAYES CLASSIFIER AND CASED BASED REASONING CLASSIFIER

Sl.No	Parameter	K-NN	Naive Bayes	Case based reasoning (lazy problem-solving)
1		Object is assigned to the class among its k nearest neighbor.	Probability is calculated and assign to class with the highest probability	Assign case which is similar to a new problem.
2	Dataset	It is applicable when dataset is small	Applicable when dataset is very large	Applicable on complicated as well as incomplete cases.
3	Performance: Increasing number of features [2][4]	Increasing number of features leads to a drop of performance.	It is perform better then k-NN with the increasing number of features	Its performance is depended on the past cases.
4	Performance: Increasing number of documents[2][4]	It perform same as Naive Bayes	Performance is almost same.	The reasoner does his/her evaluation based on what worked in the past.
5	Processing time Increasing number of documents where number of features fixed	It depends upon the size of the test data. It takes longer time to process as compare to Naive Bayes.	It also depends upon the size of the test data. Compare to k-NN Naive Bayes perform better.	processing time is fast as it provide solutions to problems Rapidly, escaping the time necessary to derive those answers from scrape.
6.	Processing time Increasing number of features where number of documents are fixed [2][4]	It takes slightly more processing time.	It take less processing time compare to k-NN	It takes less time for processing compare to both.
7	Accuracy [2]	It provides accuracy of around 72%.	It Provides accuracy of about 85%.	It provides accuracy of about 92%.

4. CONCLUSION

The Internet has changed radically the way we live, stirring communications among people at a virtual level in several environments traversing from the professional life to social relationships. The IoT has the potential to add a new aspect to this process by allowing communications with and among smart objects. Among three algorithm that have been discussed case base reasoning is better according to study as case base reasoning reduces the computational time and solve more complicated cases and also work with the huge data as well as on minimal set of solved cases furnishing the case base. This can also apply to an incomplete model. It is also easy to maintain as it can adapt many changes in problem domain by acquiring the new cases, this eliminates some need for maintenance therefore there is only the case base(s) needs to be maintained.

5. REFERENCES

[1] Chang, C.L, 1974 Finding Prototypes for Nearest Neighbour Classifiers' IEEE Trans. On Computers C-23 (11), 1179-1184

[2] Hardik Maniya Mosin I, Hasan Komal, P. Patel ,2011 Comparative study of Naive Bayes Classifier and KNN for Tuberculosis International Conference on Web Services Computing.

[3] Janet L. Kolodner , 1992 An Introduction to Case-Based Reasoning' Artificial Intelligence Review 6, 3--34, Reza Entezari-Maleki

[4] Tamije Selvy P, Palanisamy V, Elakkiya S , April 2013 Evaluation Of Classification Algorithms For Disease Diagnosis Journal of Global Research in Computer Science. Volume 4, No. 4.

[5] Website K-NN Algorithm dated 12/10/2015. https://en.wikipedia.org/wiki/K-nearest_neighbors_ algorithm.