IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

Improved KNN and other Supervised Machine Learning Approaches based on Discretization

Kalpana Kushwaha[1], Amit Thakur[2]

M. Tech Research Scholar , Department of Computer Science & Engineering, Swami Vivekanand College of Science and Technology, Bhopal[1]

Assistant Professor, Department of Computer Science & Engineering, Swami Vivekanand College of Science and Technology, Bhopal[2]

Abstract: Supervised machine learning approaches have been widely used in many applications. In this paper approaches namely Naïve Bayes, Bayesian Net, J48, Random Forest and KNN have been discussed. These algorithms are tested for sample dataset which is present in raw form. Then normalization is performed on this dataset. Precision and Accuracy has been compared for all these algorithms before and after normalization. Implementation of all these algorithms has been done using Weka. Weka is an open source tool for machine learning.

Keywords: Supervised machine learning, normalization, Weka.

I. INTRODUCTION

Data mining is that the extraction of data from great amount of data-based data sets, to find unexpected relationship and pattern hidden in information, summarize the info in novel ways in which to form it apprehensible and helpful to the info users [1,2]. Internet usage mining is that the application of data mining technique to mechanically discover and extract helpful information from a selected computer [2,3]. The term internet mining was believed to own initial came to be in 1996 by Etzioni in his paper titled ‘‘The World Wide Web: peat bog or Gold mine’’ and since then attention of analyzers world over has been shifted to the current vital research space [2,6]. In recent years, there has been AN explosive growth within the range of researches within the space of internet mining, specifically of internet usage mining. Consistent with Federico and Pier [7], over four hundred papers are printed on internet mining since the first paper printed in Nineteen Nineties.

The extremely easy Syndication (RSS) reader web site was developed for the aim of reading dailies news on-line across the world, however lack ways in which of distinguishing consumer navigation pattern and can't give satisfactory period response to the consumer desires, so, finding the suitable news becomes time overwhelming that makes the advantage of on-line services to become restricted. The study aimed toward planning and developing an automatic, online, period internet usage data processing and recommendation system supported information mercantile establishment technology. The system is in a position to observe users/clients navigation behavior by acting upon the user’s click stream information on the RSS reader computer, therefore on suggest a singular set of objects that satisfies the requirement of a full of life user in a very period, on-line basis. The user access and navigation pattern model area unit extracted from the historical access information recorded within the user’s RSS address URL file, victimization acceptable data processing techniques.

Figure 1.1 Data mining lifecycle

The K-Nearest Neighbor classification technique was used on-line and in Real- Time to use internet usage data processing technique to spot clients/visitors clicks stream information matching it to a selected user cluster and suggests a tailored browsing possibility that meet the requirement of the particular user at a given time [4]. for example, if a user looks to be sorting out politics news on china daily on his her visit to the RSS reader website, a lot of politics news headlines from different dailies like CNN politics news are going to be suggested to the user with the specified feed required to be intercalary to his/her profile so as to access such news headlines asides his/her

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

originally requested news. this can be aimed toward helping the user to urge relevant data while not expressly requesting it, therefore on ease and fasten navigation on the positioning while not too several selections being given to the user at a time, More so, the study can assist web designer and administrator to re-arrange the content of the online so as to enhance the grandness of the online site by providing online period recommendation to the consumer.Below may be a transient summary of a number of the info mining techniques consistent with completely different students within the field because it relates to our work.

II. LITERATURE REVIEW

Bayes Classifier— It originates from previous works in pattern recognition and is linked to the family of probabilistic Graphical Models. For each class, a probabilistic summary is stored. The conditional probability of each attribute and the probability of the class are stored in this summary. The graphical models are used to display knowledge about domains which are uncertain in nature. In the graphs [15], nodes depict random variables and the edges which connect corresponding random variable nodes are assigned weights which represent probabilistic dependencies. On encountering a new instance, the algorithm just creates an update of the probabilities stored along with the specific class [12]. The sequence of training instances and the existence of classification errors do not have any role in this process. Thus basically it has to predict the class depending on the value of the members of the class. This category consists of 13 classifiers, but only 3 of those are compatible with our chosen dataset.

Function classifier— It deploys the concept of regression and neural network. Input data is mapped to the output. It employs the iterative parameter estimation scheme. Overall there are 18 classifiers under this category, out of which only 2 are compatible with our dataset.

J48— It is an enhanced version of C 4.5 which revolves on the ID3 algorithm with some extra functionalities to resolve issues that ID3 was incompetent in [10]. However, this technique is time and space consuming. Initially, it builds a tree using the divide and conquer algorithm and then applies heuristic criteria. The rules according to which the tree is generated are precise and intuitive [2].

COMPARISON OF CLASSIFIERS			
CLASSIFIER	CATEGORY	DESCRIPTION	REFERENCE
Naive Bayes	Probability based classifier	This is a probability based classifier based on Naive Bayes conditional probability	[15]
Bayesian Net	Probability based classifier	This is a probability based classifier based on Naive Bayes conditional probability.	
J48	Tree based approach	It is enhanced version of C 4.5 algorithm and used ID3.	[25]
Random Forest	Tree based approach	It is also a decision tree based approach but have more accuracy as compared to J48.	[25]
Random Tree	Tree based approach	It generates a tree by randomly selecting branches from a possible set of trees.	[25]
REPTree	Tree based approach	It uses gain and variance for prediction.	[17]

IV. RESULTS

1) KNN Algorithm

BEFORE NORMALIZATION			
TP RATE	**FP RATE**	**PRECISION**	**ACCURACY**
0.690	0.531	0.685	69%
AFTER NORMALIZATION			
TP RATE	**FP RATE**	**PRECISION**	**ACCURACY**
0.993	0.007	0.993	99.33%

2) NAÏVE BAYES

BEFORE NORMALIZATION			
TP RATE	**FP RATE**	**PRECISION**	**ACCURACY**
0.862	0.095	0.893	86.17%
AFTER NORMALIZATION			
TP RATE	**FP RATE**	**PRECISION**	**ACCURACY**
0.980	0.020	0.981	98%

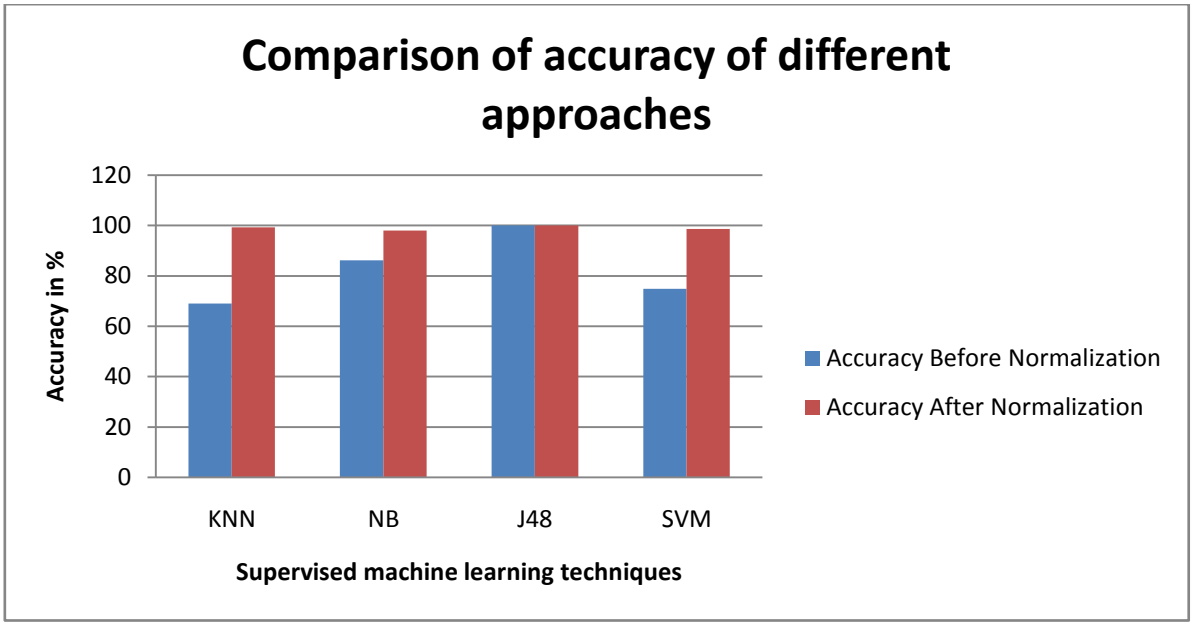
3) J48

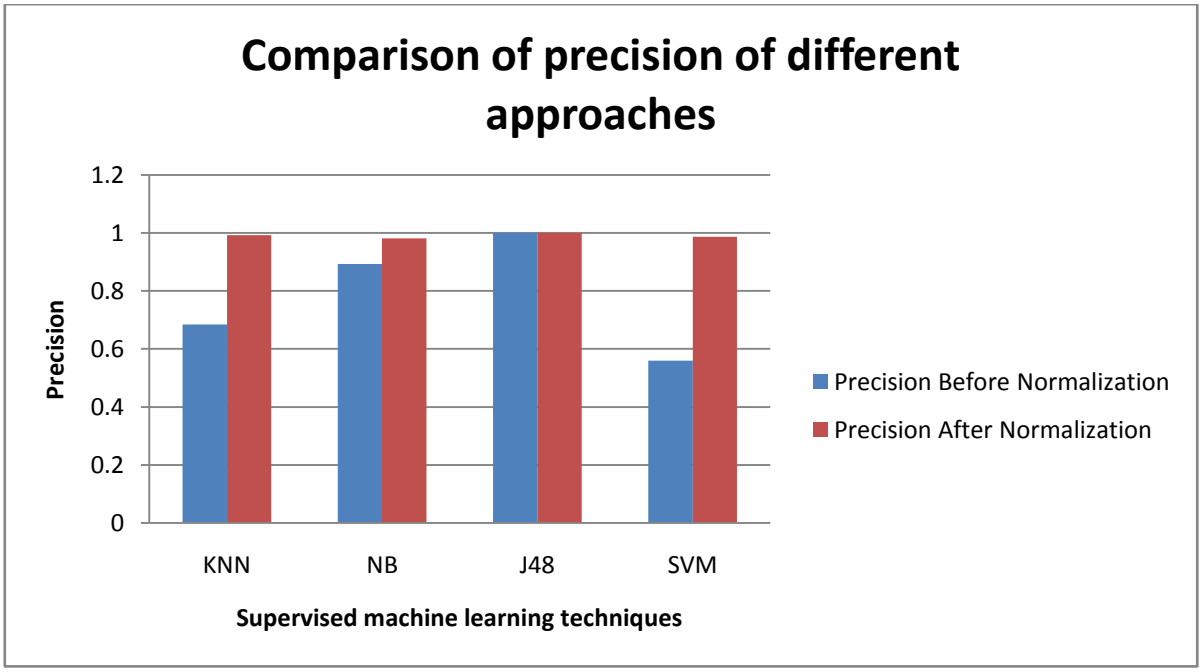
BEFORE NORMALIZATION			
TP RATE	**FP RATE**	**PRECISION**	**ACCURACY**
1	0	1	100%
AFTER NORMALIZATION			
TP RATE	**FP RATE**	**PRECISION**	**ACCURACY**
1	0	1	100%

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

4) SVM

BEFORE NORMALIZATION			
TP RATE	**FP RATE**	**PRECISION**	**ACCURACY**
0.748	0.748	0.560	74.83%
AFTER NORMALIZATION			
TP RATE	**FP RATE**	**PRECISION**	**ACCURACY**
0.987	0.009	0.987	98.67%





V. CONCLUSION

In this era of data analytics, machine learning has emerged as a vital domain of research. For classification and clustering of datasets different machine learning techniques have been Machine learning algorithms can be broadly classified as supervised and unsupervised approaches. In this dissertation supervised machine learning techniques which include KNN, Naïve Bayes, Support vector machine and J48 have been studied. But

the objective of research is to increase the accuracy of classification for the raw datasets. Here normalization has been applied on the raw dataset and it is found that accuracy has been improved after supervised discretization of dataset.

In future the proposed scheme can be tested for other datasets also. And normalization and feature selection approaches can be studied and suitable can be applied.

REFERENCES

[1] T. Luigi, S. Giacomo, Mining frequent item sets in data streams within a time horizon, J. Data Knowledge Eng. 89 (2014) 21–37.
[2] K. Mi-Yeon, H.L. Dong, Data-mining based SQL injection attack detection using internal query trees, J. Expert Syst. Appl. 41 (2014) 5416–5430.
[3] C. Luca, G. Paolo, Improving classification models with taxonomy information, J. Data Knowledge Eng. 86 (2013) (2013) 85–101.
[4] A. Dario, B. Eleno, B. Giulia, C. Tania, C. Silvia, M. Naeem, Analysis of diabetic patients through their examination history, J. Expert Syst. Appl. 40 (2013) 4672–4678.
[5] F.N. David, Data mining of social networks represented as graphs, J. Comput. Sci. Rev. 7 (2013).
[6] L. Shu-Hsien, C. Pei-Hui, H. Pei-Yuan, Data mining techniques and applications- A decade review from 2000 to 2011, Journal of expert system with applications 39 (2012)) 11303–11311.
[7] M. Michal, K. Jozef, S. Peter, Data preprocessing evaluation for web log mining: reconstruction of activities of a web visitor, J. Proc. Comput. Sci. 1 (2012) 2273–2280.
[8] A. Niyat, K. Amit, K. Harsh, A. Veishai, Analysis the effect of data mining techniques on database, Journal of advances in Engineering & software 47 (2012) 164–169.
[9] T. Rivas, M. Paz, J.E. Martins, J.M. Matias, J.F. Gracia, J. Taboadas, Explaining and predicting workplace accidents using data-mining Techniques, Journal of Reliable Engineering and System safety 96 (7) (2011) 739–747.
[10] MySQL Corporation, MySQL Database Management System Software. USA MySQL/Oracle Corporation, 2008.
[11] NetBeans IDE 7.3, NetBeans java compiler. USA, Java/Oracle corporation, 2008.
[12] I.O. Ogbonaya, Introduction to Matlab/Simulink, for engineers and scientist, 2nd edition., John Jacob's Classic Publishers Ltd, Enugu, Nigeria, 2008.
[13] D. Resul, T. Ibrahim, Creating meaningful data from web log for improving the impressiveness of a web site by using path analysis method, Journal of expert system with applications 36 (2008) 6635–6644.
[14] C. Padraig, J.D. Sarah, K-Nearest Neighbor Classifier. Technical Report UCD-CSI-2007-4, University College Dublin, 2007.
[15] S. Amartya, K.D. Kundan, Application of Data mining Techniques in Bioinformatics, B.Tech Computer Science Engineering thesis, National Institute of Technology, (Deemed University), Rourkela, 2007.
[16] Z. Xuejuu, E. John, H. Jenny, Personalised online sales using web usage data mining, J. Comput. Ind. 58 (2007) 772–782.
[17] M. Zdravko, T.L. Daniel, Data mining the Web, Uncovering patterns in Web content, structure, and usage, John Wiley & sons Inc., New Jersey, USA, 2007, p. 115–132.
[18] Z. Xuejuu, E. John, H. Jenny, Personalised online sales using web usage data mining, J. Comput. Ind. 58 (2007) 772–782.
[19] L. Habin, K. Vlado, Combining mining of web server logs and web content for classifying users' navigation pattern and predicting users future request, J. Data Knowledge Eng. 61 (2007).
[20] M. Zdravko, T.L. Daniel, Data mining the Web, Uncovering patterns in Web content, structure, and usage, John Wiley & sons Inc., New Jersey, USA, 2007, p. 115–132.