International Workshop on Data Mining in IoT Systems (DaMIS 2016)

Analysis of Eight Data Mining Algorithms for Smarter Internet of Things (IoT)

Furqan Alam\textsuperscript{a}, Rashid Mehmood\textsuperscript{b,\ast}, Iyad Katib\textsuperscript{a}, Aiiad Albeshri\textsuperscript{a}

\textsuperscript{a}Department of Computer Science, Faculty of Computing and Information Technology (FCIT), King Abdulaziz University, Jeddah, KSA
\textsuperscript{b}High Performance Computing Center, King Abdulaziz University, Jeddah, KSA

Abstract

Internet of Things (IoT) is set to revolutionize all aspects of our lives. The number of objects connected to IoT is expected to reach 50 billion by 2020, giving rise to an enormous amount of valuable data. The data collected from the IoT devices will be used to understand and control complex environments around us, enabling better decision making, greater automation, higher efficiencies, productivity, accuracy, and wealth generation. Data mining and other artificial intelligence methods would play a critical role in creating smarter IoTs, albeit with many challenges. In this paper, we examine the applicability of eight well-known data mining algorithms for IoT data. These include, among others, the deep learning artificial neural networks (DLANNs), which build a feed forward multi-layer artificial neural network (ANN) for modelling high-level data abstractions. Our preliminary results on three real IoT datasets show that C4.5 and C5.0 have better accuracy, are memory efficient and have relatively higher processing speeds. ANNs and DLANNs can provide highly accurate results but are computationally expensive.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). Peer-review under responsibility of the Program Chairs

Keywords: Internet of Things (IoT), Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Linear Discriminant Analysis (LDA), Naive Bayes (NB), C4.5, C5.0, Artificial Neural Networks (ANNs), Deep Learning ANNs (DLANNs), Big Data, Smart Cities;

1. Introduction

The Internet of Things (IoT) is “a global infrastructure for the information society, enabling advanced services by interconnecting (physical and virtual) things based on existing and evolving interoperable information and communication technologies”\textsuperscript{1}. IoT\textsuperscript{2,3}, is one of the major technological developments of our times given its potential is fully realized. It is set to revolutionize all aspects of our lives, be it work, social interactions, or entertainment. A plethora of objects are emerging every day to be part of the IoT infrastructure and the number of objects connected to IoT is expected to reach 50 billion by 2020\textsuperscript{4}. This would generate enormous amounts of valuable data.

A major objective of IoT is to make the environment around us smarter, by giving the environment the information it needs, through real-time and historic data feeds, and apply computational intelligence on the information to make

\textsuperscript{\ast} Corresponding author. Tel.: +966-537-069-318. 
\textit{E-mail address:} RMehmood@kau.edu.sa
smart decisions automatically. This would enhance our ability to manage our cities, roads, health, homes, forests and much more. We will have better crowd and traffic management, emergency predictions, better prediction of accidents and crimes, etc. The data collected from IoT devices will be used to understand and control complex environments around us, enabling better decision making, greater automation, higher efficiencies, productivity, accuracy, and wealth generation. A major challenge in these settings is the timely analyses of large amounts of data (i.e. big data) to produce highly reliable and accurate insights and decisions so that IoT could live up to its promise. Machine learning is among the top methods to gain hidden insights from IoT data.

The aim of this research is to explore whether the conventional data mining algorithms would also work for the IoT datasets, or new families of data mining algorithms are required. To this end, this paper provides a preliminary analysis on examining the applicability of several well-known data mining algorithms to real IoT datasets. We have used eight data mining algorithms. These are Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Linear Discriminant Analysis (LDA), Naïve Bayes (NB), C4.5, C5.0, Artificial Neural Networks (ANNs), and Deep Learning ANNs (DLANNs). The main contribution of this work is the analysis of the effectiveness and efficiency of eight of the well-known data mining algorithms.

The remainder of this paper is organized as follows. Section 2 provides a review of the relevant literature. Our experimental methodology is explained in Section 3. Simulation results and analysis are reported in Section 4. Conclusions are drawn in Section 5.

2. Background Material and Literature Review

Modern day data mining tasks are far more challenging due to an unprecedented increase in the amount and complexity of data. With the emergence of IoT paradigm, a completely new set of challenges have been added to the data mining domain. Support Vector Machine (SVM), K-nearest neighbor (KNN), Naïve Bayes (NB), Linear discriminant analysis (LDA), C4.5, C50 and ANNs are widely used in the field of data mining. SVM was designed initially to address bias variance tradeoff, over-fitting and capacity control. Burges also stated that SVM accuracy depends a lot on the quality of training data and machine capacity. The use of SVM is further extended from classification to regression and element ranking. SVM is a very efficient tool to work with in complex and noisy domains. The Computational inefficiency is one of the major drawback of SVM, however several optimizations have been done to reduce its computational cost and to increase its scalability.

A lazy learner algorithm known as KNN is one of the simplest available classifier which is easy to understand and implement. Due to the simplicity of KNN, several issues arise that limit its performance such as the selection of right distance measure, number of neighbors and majority vote to combing class labels is not always effective. KNN is also used effectively for various tasks in wireless sensor networks (WSN) and IoT domain for intrusion detection. LDA which is also known as Fisher Discriminant Analysis (FDA), is an appropriate answer for multi-class problems. The C4.5 algorithm, another popular data mining algorithm, proposed by Ross Quinlan is considered as the best algorithm in data mining. Quinlan later on developed an improvised version of C4.5, known as C5.0, which is claimed faster than C4.5, has better memory efficiency, and support for boosting, winnowing and weighting.

The approach taken by the NB and ANN algorithms to solve the given data mining task is completely different to the ones we have discussed earlier in this section. The NB algorithm is probability based and is a very old classifier. It uses Bayes’ theorem with the independence supposition among the features. NB is a robust and simple classifier like KNN. Although simple, it often produces surprisingly accurate results. On the other side, ANN algorithms are based on mimicking neural system of human brain. ANNs are extremely efficient in solving data mining tasks with higher accuracy. However, ANN based algorithms are too complex and an extensive amount of computation is required for solving a problem with higher accuracy. Further extension of ANNs are the revolutionary learning algorithms based on deep learning concept. DLANNs have extreme learning ability, process enormous amount of data, and produce highly accurate results which are not possible with other conventional machine learning and data mining algorithms. However, this deep learning branch of computer science is still in its infancy period. Deep learning can give us novel insights from IoT data which is not possible from other data mining algorithms. Particularly, with respect to IoT, little work is done to gain benefits from DLANNs.
3. Experimental Methodology

We have considered eight well-known data mining algorithms in this paper. These eight algorithms also include DLANNs which build feed-forward multilayer ANNs. All the simulations are performed using the R platform. Particularly, for simulating DLANNs, we used the H2O package available in R. For the experiments, we used three real sensor datasets from the UCI data repository. Datasets are collected by using sensors and accelerometers and are used to classify human activities, robot navigation, body postures and movements. Before simulating the algorithms, we preprocessed the datasets to make them suitable for the classifiers. This is a preliminary analysis and hence, we have only used partial datasets. Our experimental methodology is depicted in Fig. 1.

4. Results and Analysis

We now present a preliminary analysis of the eight well established data mining algorithms as mentioned in Fig. 1. The experiments have been carried out on the Aziz supercomputer. The Aziz supercomputer is Fujitsu made and is able to deliver peak performance of 230 teraflops. It has a total of 11,904 cores in 496 nodes. Aziz was ranked number 360 in the June 2015 Top500 competition (http://www.top500.org/), currently it is at number 491 (November 2015). For performance evaluation of the algorithms, we have summarised the results in form of confusion matrix (CM). With the help of CM, we are able to know the total number of instances rightly and wrongly classified. The class to which the wrongly classified instances belong to can also be identified with CM.

The CMs of all eight algorithms simulated on three different datasets are given in Fig. 2 to Fig. 4. Moreover, Table 1, classification accuracy (CA) percentage and elapsed time is mentioned. With respect to Fig. 2 to Fig. 4 and Table 1, we note that C4.5, C5.0, ANN and DLANN algorithms performed far better than SVM, KNN, NB and LDA. The C4.5, C5.0 and ANN algorithms are very close to each other when considering the classification accuracy. With average accuracy (AAC) of 97.15% obtained by C4.5, it performs slightly better than 96.61% AAC of the C5.0 algorithm. AAC of ANN is 96.19% for all three datasets. C4.5 tops among all the eight algorithms in terms of the classification accuracy, followed closely by C5.0.

All the datasets are multi-label as mentioned in Fig. 2 to Fig. 4. Consequently, SVM shows its weaknesses towards multi-label data classification as compared to binary classification problem where its performance is one of the best. SVM performed better than KNN with 4.09% higher AAC. The choice of k-neighbours and distance measure affects the CA of KNN. The remaining two algorithms, NB and LDA, performed the worst in terms of CA. Our these results strongly agree to the conclusion made by Wu et al.

4.1. Execution Time

Both NB and LDA algorithms are the fastest amongst all the eight algorithms. LDA has slightly better processing times than NB. LDA will not have higher CA if discriminatory information is in the variance. Also, it is a parametric method. Average processing time (APT) of C4.5, C5.0 is 7.70 and 7.21 seconds respectively. SVM uses a lot of system resources and has slow processing speed. KNN is lighter and has low execution times as mentioned in Table 1.
### Fig. 2. Confusion matrix of (a) SVM; (b) KNN; (c) NB; (d) C4.5, (e) LDA; (f) C5.0; (g) ANNs and (h) DLANNs for dataset 29

<table>
<thead>
<tr>
<th>Actual/Predicted</th>
<th>Sitting</th>
<th>Sitting Down</th>
<th>Standing</th>
<th>Standing Up</th>
<th>Walking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sitting</td>
<td>50</td>
<td>2</td>
<td>1</td>
<td>33</td>
<td>1</td>
</tr>
<tr>
<td>Sitting Down</td>
<td>2</td>
<td>115</td>
<td>139</td>
<td>103</td>
<td>50</td>
</tr>
<tr>
<td>Standing</td>
<td>0</td>
<td>3</td>
<td>471</td>
<td>127</td>
<td>82</td>
</tr>
<tr>
<td>Standing Up</td>
<td>48</td>
<td>260</td>
<td>267</td>
<td>11,806</td>
<td>34</td>
</tr>
<tr>
<td>Walking</td>
<td>106</td>
<td>979</td>
<td>85</td>
<td>4,292</td>
<td>1</td>
</tr>
</tbody>
</table>

### Fig. 3. Confusion matrix of (a) SVM; (b) KNN; (c) NB; (d) C4.5, (e) LDA; (f) C5.0; (g) ANNs and (h) DLANNs for dataset 30

<table>
<thead>
<tr>
<th>Actual/Predicted</th>
<th>Sitting</th>
<th>Sitting Down</th>
<th>Standing</th>
<th>Standing Up</th>
<th>Walking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sitting</td>
<td>25</td>
<td>9</td>
<td>1</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>Sitting Down</td>
<td>1</td>
<td>589</td>
<td>48</td>
<td>59</td>
<td>16</td>
</tr>
<tr>
<td>Standing</td>
<td>0</td>
<td>4</td>
<td>247</td>
<td>44</td>
<td>36</td>
</tr>
<tr>
<td>Standing Up</td>
<td>14</td>
<td>106</td>
<td>93</td>
<td>597</td>
<td>23</td>
</tr>
<tr>
<td>Walking</td>
<td>5</td>
<td>67</td>
<td>280</td>
<td>55</td>
<td>215</td>
</tr>
</tbody>
</table>

Fig. 2. Confusion matrix of (a) SVM; (b) KNN; (c) NB; (d) C4.5, (e) LDA; (f) C5.0; (g) ANNs and (h) DLANNs for dataset 29

<table>
<thead>
<tr>
<th>Actual/Predicted</th>
<th>Sitting</th>
<th>Sitting Down</th>
<th>Standing</th>
<th>Standing Up</th>
<th>Walking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sitting</td>
<td>50</td>
<td>35</td>
<td>2</td>
<td>77</td>
<td>2</td>
</tr>
<tr>
<td>Sitting Down</td>
<td>1215</td>
<td>613</td>
<td>2574</td>
<td>1403</td>
<td>504</td>
</tr>
<tr>
<td>Standing</td>
<td>353</td>
<td>43105</td>
<td>61</td>
<td>3651</td>
<td></td>
</tr>
<tr>
<td>Standing Up</td>
<td>1448</td>
<td>2002</td>
<td>2789</td>
<td>5579</td>
<td>597</td>
</tr>
<tr>
<td>Walking</td>
<td>59</td>
<td>209</td>
<td>12254</td>
<td>834</td>
<td>8038</td>
</tr>
</tbody>
</table>

Fig. 3. Confusion matrix of (a) SVM; (b) KNN; (c) NB; (d) C4.5, (e) LDA; (f) C5.0; (g) ANNs and (h) DLANNs for dataset 30
ANNs and DLANNs have higher computational requirements. For IoT, there can be problems where high CA does not matter much but processing time matters; in those cases, NB and LDA can be handy.

4.2. The Deep Learning Algorithm

Based on our preliminary assessment, we believe that DLANNs can have the best CA among all the simulated algorithms. We observed that an improved classification accuracy could be achieved by increasing the epochs, hidden layers and neurons. In DLANNs, CA also depends significantly on its parameters tuning. DLANNs have a very complex structure, need a large amount of system resources, and as a result, DLANN algorithm has the highest execution time among all the eight algorithms presented in this work.

Table 1. Classification accuracy in % and elapsed time in seconds for all the algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Dataset 29</th>
<th>Dataset 30</th>
<th>Dataset 31</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>98.57</td>
<td>86.43</td>
<td>91.75</td>
</tr>
<tr>
<td>KNN</td>
<td>98.94</td>
<td>86.98</td>
<td>91.75</td>
</tr>
<tr>
<td>NB</td>
<td>77.04</td>
<td>70.09</td>
<td>91.75</td>
</tr>
<tr>
<td>C4.5</td>
<td>99.69</td>
<td>91.81</td>
<td>91.75</td>
</tr>
<tr>
<td>C5.0</td>
<td>99.62</td>
<td>90.26</td>
<td>91.75</td>
</tr>
<tr>
<td>LDA</td>
<td>81.85</td>
<td>71.53</td>
<td>66.45</td>
</tr>
<tr>
<td>ANN</td>
<td>99.03</td>
<td>89.55</td>
<td>100.55</td>
</tr>
<tr>
<td>DLANN</td>
<td>99.52</td>
<td>87.10</td>
<td>98.49</td>
</tr>
</tbody>
</table>

5. Conclusion

The IoT paradigm brings new sets of data mainly collected from sensor devices. To capture this hidden knowledge from IoT data is a challenging task in data mining. Some researchers argue that a new family of data mining algorithms are needed to handle IoT data. In our work, we examined the applicability of some of the well established data mining algorithms like SVM, KNN, NB, C4.5, LDA, C5.0, ANNs and DLANNs for dataset 31.
algorithms including DLANNs. With our preliminary analysis, we conclude that C4.5, C5.0, ANNs and DLANNs can give relatively higher accuracy results. We plan to conduct a detailed study on larger and diverse IoT datasets in the future.

Acknowledgements

The experiments reported in this paper were performed on the Aziz supercomputer at King AbdulAziz University, Jeddah, Saudi Arabia. We are also thankful to UCI data repository for making datasets available.

References