ISSN 2278-3091
Volume 8, No.1.1, 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse4481.12019.pdf
https://doi.org/10.30534/ijatcse/2019/4481.12019

Agent-based Big Data Mining

Nojod M. Alotaibi[1], Manal Abdullah[2], Hala Mosli[3]
[1,2]Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University Jeddah, Saudi Arabia
[1]nojodalotaibi@gmail.com
[2]maaabdullah@kau.edu.sa
[3]Endocrinology and Metabolism Department of Medicine, King Abdulaziz University Jeddah, Saudi Arabia
[3]hmosli@kau.edu.sa

ABSTRACT

Big data is the term used to describe a data, which is difficult to process, manage and analyze using traditional databases or data mining algorithms. Useful knowledge can be extracted from this big data with the help of data mining. Due to the volume, variety, and velocity of data, traditional techniques of data mining may be unsuitable to work with big data. As a result, there is a basic need to create powerful and productive enormous information mining methods. Characterization is one of the information mining strategies that is capable of processing a large amount of data and is growing in popularity. It is used to map a data item into one of several predefined classes or categories. Healthcare data is one form of big data not only for its sheer volume, but also for its complexity, diversity, and speed at which it is generated and must be managed. In this paper, we present the problem of mining the big data using software agent. The main goal of this paper is to develop and implement an agent-based big data classification model that can predict the severity of diabetes disease. Results proved that using agent technology in the preprocessing stage saved the memory storage from 8.66 TB to 5 GB memory space. The transfer data time is reduced from about 12 days to about 10 minutes after preprocessing data remotely using the agent. Regarding classification accuracy, the proposed model has proven 87% accuracy and 65% reliability.

Key words—Big Data; Data Mining; Classification; Agent; Random Forest.

1. INTRODUCTION

Knowledge discovery is the overall process for extracting relevant and interesting knowledge from data, while data mining is a particular step in this process. There are many techniques of data mining used for different purposes and goals. Classification is one of the most important data mining techniques. It is used for classifying data into different classes. Several classification algorithms have been widely used in various domains such as: decision trees, support vector machines, genetic algorithms, neural networks, etc. However, dealing with a large sample size and high dimensional data may cause several issues with respect to scalability and learning performance of these classification algorithms. Moreover, the classification ability of a single classifier is limited. One of the solutions to this problem is to train a group of classifiers on these large datasets instead of a single classifier. The technique of using multiple classifiers for solving the same problem is known as ensemble learning. The Ensemble classification system consists of a set of individual classifiers (called base/weak classifiers) that are trained in different aspects of the problem and then, the outputs of these classifiers are combined together using a majority vote to produce the ensemble output [1]. The main goal of the ensemble is to achieve higher classification accuracy than that produced by its members (the individual classifiers that make it up). Examples of ensemble classifiers are bagging, boosting, and random forest [1]. Random forest is a gathering arrangement technique comprising of different un-pruned choice trees. It frames bootstrap tests by haphazardly apportioning the first component space as opposed to utilizing the entire info highlights [2]. At the hub determination arrange, it chooses the best part hub from an arbitrarily chosen set of hubs. Then, the predictions from all trees are combined through a majority vote to generate the final prediction.

Smart specialists are another worldview for creating programming applications. As of late, operators have been utilized to improve and bolster the information revelation process. For example, operators can contribute in information determination, extraction, preprocessing, and coordination. They are useful in lessening the all out handling time and data over-burden, in complex undertakings, for example, medicinal checking and war zone thinking [3].

This paper is comprised of the model of agent based big data classification. Researchers used JADE as an implementation tool for agents. The classification is performed using the random forest algorithm from Apache Mahout.

The remainder of this paper is organized as follows. Section 2 presents some related works. The proposed model is detailed in Section 3. Section 4 describes the dataset. Section 5 presents the agent effect on the data transfer time and memory capacity. Experimental results provided and analyzed in Section 6. Finally, authors give some conclusions in Section 7.

2. RELATED WORKS

There are many studies that are presented in this section that discuss the classification using software agent. Among them, Song et al. [4] developed effective information agents to autonomously classify and filter documents, by simply analyzing their titles. The classification agent uses an information, inference model that allows it to discover implicit information in the document titles. The filtering agent uses the belief revision model to determine which categories can interest the user. The authors claimed that their document classification and filtering mechanism outperforms the Support Vector Machines (SVM) model. Kalaivani and Shunmuganathan [5] developed an agent based system for classifying sentiment of online reviews using ontology. The proposed system consists of three main components: data extraction agent, recommendation agent, and feature selection agent. Finally, the results from this system are used for re ranking the book's results. Pinzon et al. [6] developed a mechanism of classification based on a CBR-BDI agent. This mechanism classifies the incoming SOAP message and rejects the malicious SOAP messages by using a case-based reasoning mechanism (CBR). There are two phases in the classification process. The CBR mechanism incorporates a decision tree and fuzzy logic rules in the first phase, while incorporating neural networks in the second phase. Zolfaghar et al. [7] developed a big data framework for predicting 30-day risk of readmission for Congestive Heart Failure (CHF) patients. First, they extracted useful factors from National Inpatient Dataset (NIS) and augmented it with CHF patient's dataset from Multicare Health System (MHS) using Hive and Cassandra. Then, scalable data mining models were developed to predict the risk of readmission using Random Forest algorithm in Mahout. Ko et al. [8] developed a new methodology to profile specific medical information from patient medical records for predicting the progression of amyotrophic lateral sclerosis (ALS) disease. They implemented a system using HBase on Hadoop and the Random forest classifier of Apache Mahout. Training and testing data were generated from features of medical records provided by the Pooled Resource Open- Access ALS Clinical Trials Database (PRO-ACT) site [9].

The decision supporting system consists of these three modules: data management, generation of feature matrix, and prediction of ALS progress. Kiruba and Arasu [10] built up a canny operator based framework for Liver Disorder conclusion. This framework is assessed utilizing two liver malignant growth datasets to be specific the BUPA dataset and the Indian Liver Patient dataset (ILPD). The agents are designed using the JADE environment wherein three types of agents are required for the design of the clinical system: Diagnosis support agents (DSA), Decision agents (DA), and Learning agents (LA). The test result demonstrated that the C4.5 choice tree calculation and the Random Tree calculation created 100% exactness in arrangement of the liver issue. Kulothungan et al. [11] proposed an insightful operator based interruption recognition framework utilizing a specialist based staggered grouping for identifying the gatecrashers in remote sensor systems (WSNs). The grouping is performed by a blend of two characterization calculations, to be specific Enhanced Decision Tree classifier (C4.5) and Enhanced Multiclass Support Vector Machine (MSVM). In this work, the attention was on expanding the discovery precision and diminishing false positive rates. This framework comprises of three modules where the tree classifier specialist utilizes the Enhanced C4.5 calculation with operator choice for building a choice tree, which is utilized to discover abuse discovery. The grouping module utilizes Agent Multiclass SVM for unsupervised abnormality recognition. At last, for refined grouping of irregularity discovery, the specialist based tree classifier has been utilized in this work.

3. SYSTEM MODEL

In this section, an agent-based big data classification model is proposed. Data is gathered from distributed data sources and preprocessed by software agents. To achieve the research goal, an agent development tool "JADE" is used as a programming framework for system implementation. In addition, Hadoop as a big data framework is used to achieve performance, scalability, and fault tolerance for the task at hand. For the classification task, the research incorporates a random forest module from Apache Mahout into a model. Apache Mahout is a library of scalable machine learning and data mining algorithms that are free to use under the Apache license. It contains implementation for clustering, categorization, collaborative, filtering, and evolutionary programming on top of Hadoop and using the Map/Reduce paradigm. The proposed model consists of two main modules as depicted by Figure 1. The two modules are: an agent-based preprocessing module and a data mining module. The next subsections will detail these modules.

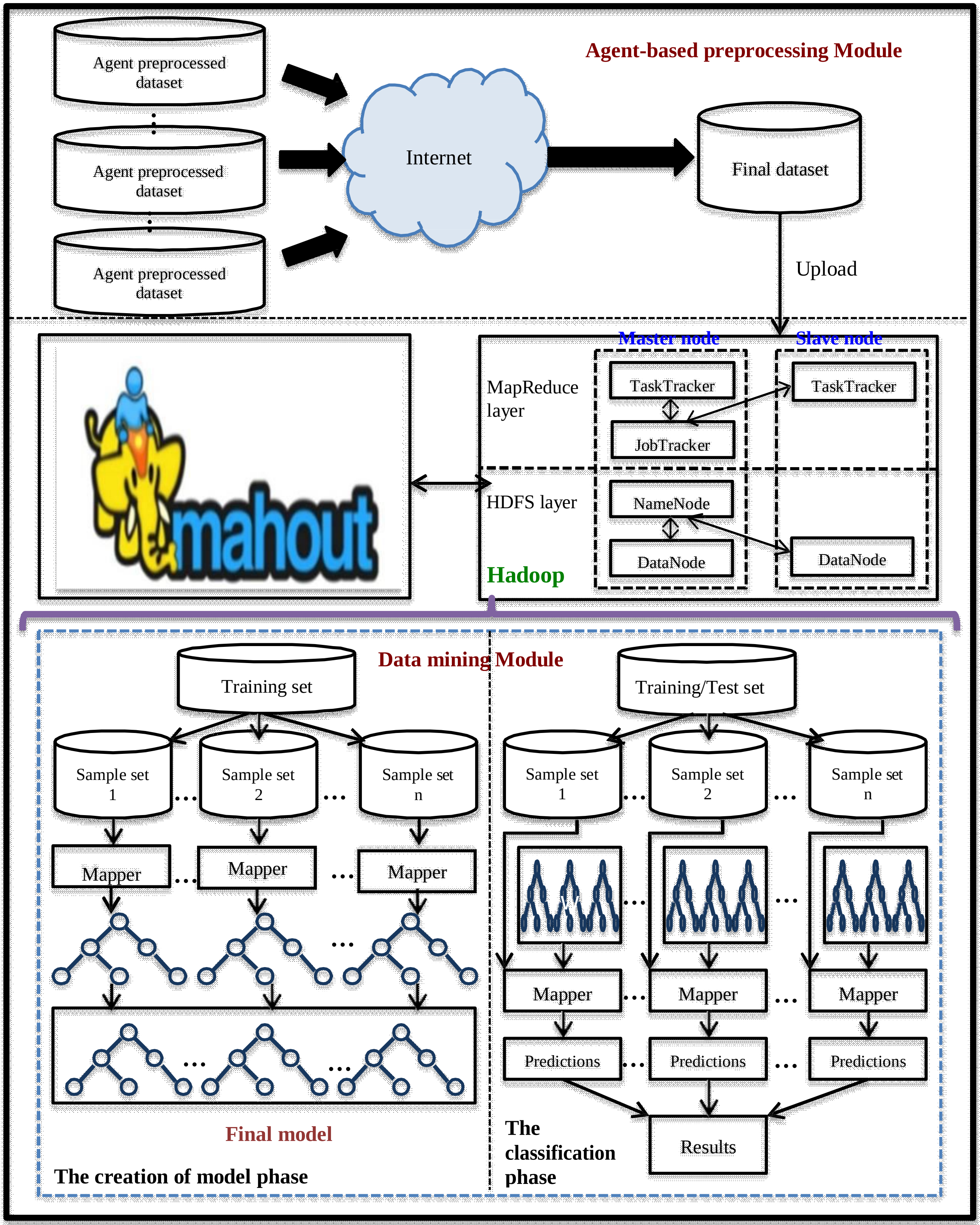


Figure 1: The overall architecture of the proposed model

A. Agent-based Preprocessing Module

In the era of big data applications, data comes in different types from different sources. However, preprocessing a large database is very time-consuming and too cumbersome. As a solution to this problem, this research proposes an agent-based preprocessing module as shown in Figure 2. The aim of this module is to provide useful assistance by reducing the time and effort required by the analyst to accomplish its task. In the agent-based preprocessing module, we have different data sources with different types and sizes that satisfy the 5 V's (Volume, Variety, Velocity, Veracity, and Value) of big data. By using an agent, authors can extract the most useful and related data to the classification process. At each data source, the agent is responsible for:

- Omitting instances with missing values from the dataset.
- Choosing the remarked features we are interested in for some disease.
- Converting unstructured data such as doctor's notes and MR (magnetic resonance) images into structured data.

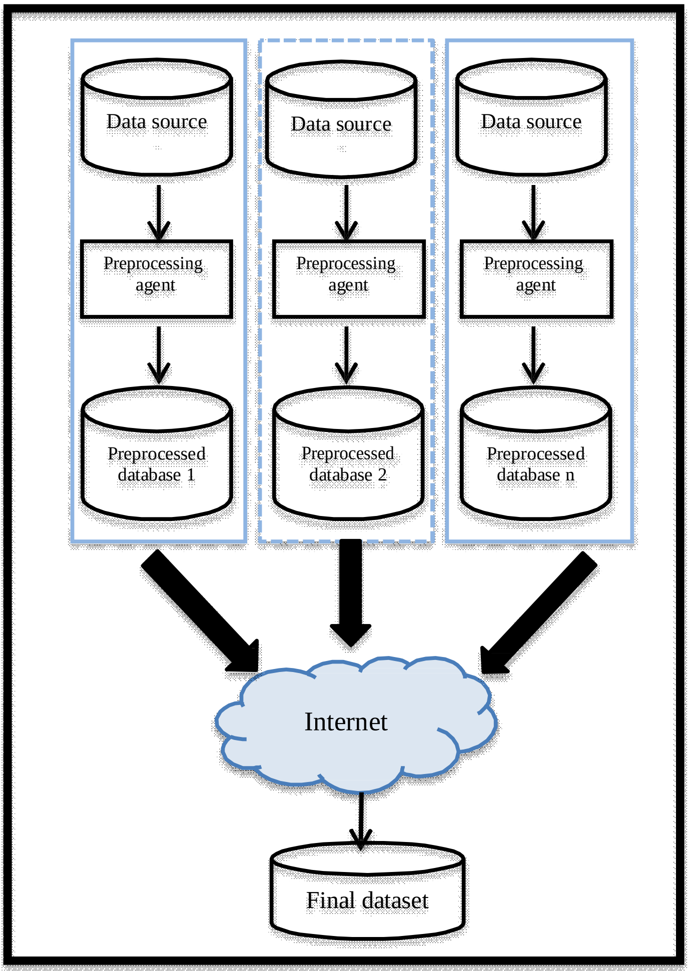


Figure 2: Agent-based preprocessing module

B. Data Mining Module

As mentioned earlier, the proposed model has two main modules: an agent-based module and a data mining module. In, the datasets coming from different data sources are preprocessed and then combined to produce the final dataset. Then, the output of this module becomes input to the data mining module. The data mining module consists of three main steps organized in the following order. Firstly, the final dataset is loaded into HDFS. Secondly, the dataset is divided into two sets: training and testing sets. Finally, the Mahout random forest algorithm is applied to the training set. The two main phases of this algorithm are summarized below.

The primary stage is the making of the model, where an irregular backwoods is worked from the first preparing set after a Map/Reduce method. In this stage, the preparation set is fragmented into autonomous information squares; at that point, these squares are reproduced and exchanged between the distinctive handling hubs. Next, each Map errand assembles a subset of the woodland (a few arbitrary trees of the timberland) with the information square of its segment and creates a record containing the manufactured trees. At long last, the yield documents created by the mappers are parsed to extricate the trees. The gathering of all trees frames the backwoods. The second stage is the estimation of the classes related with the dataset utilizing the recently learned model. In this stage, the dataset (preparing or testing set) is sectioned into autonomous information squares; recreates and exchanges them to different machines to be at long last handled freely by each guide errand in parallel. Next, every mapper assesses the class for the cases accessible in it utilizing a dominant part vote of the anticipated class by the trees in the irregular woodland show worked in the past stage. At last, the expectations produced by every mapper are connected to frame the last forecasts record.

4. DATASET DESCRIPTION

The dataset utilized in this examination is from the Cerner Health Facts database (Cerner Corporation, Kansas City, MO), a national information stockroom that gathers extensive clinical records crosswise over medical clinics all through the United States. This database contains information deliberately and persistently gathered from partaking establishments electronic restorative records and involves experience information (crisis, outpatient, and inpatient), supplier forte, socioeconomics (age, sex, and race), analyze and in-emergency clinic methods archived by ICD-9-CM codes, research facility information, drug store information, in-medical clinic mortality, and medical clinic attributes [12]. The information speaks to 10 years (1999-2008) of clinical consideration at 130 emergency clinics and coordinated conveyance organizes all through the United States. The first database comprises of 41 tables in a reality measurement outline and an aggregate of 117 highlights. It incorporates 74,036,643 one of a kind experiences (visits) that compare to 17,880,231 one of a kind patients and 2,889,571 suppliers [12]. The dataset used was taken out from the database for experiences that fulfilled the accompanying criteria [12]:

1- It is an inpatient experience (a medical clinic confirmation).
2- It is a "diabetic" experience, that is, one amid which any sort of diabetes was entered to the framework as a finding.
3- The length of stay was no less than 1 day and at most 14 days.
4- Laboratory tests were performed amid the experience.
5-Medications were directed amid the experience.

The initial dataset includes 101,766 instances and 55 features representing patient and hospital outcomes. This is available from the UCI Machine Learning Repository as the "Diabetes 130-US hospitals for years 1999-2008 Data Set" [13]. Some data preprocessing is required to align the data with the aim of this study. This is the responsibility of the agent module to preprocess data as depicted in the system model in Figure 2. First, removing all instances with a missing class value from the dataset. Also removed several features that are irrelevant to our goal such as admission type, admission source, discharge disposition, etc.

The dataset has 24 features for oral diabetes medications with values "No", "Up", "Down", and "Steady". We combined the "Up", "Down" and "Steady" into a single value called "Yes".

However, the oral medications used to treat diabetes are commonly classified into several groups or classes based on their mechanism of action. The patient may take medicines from one or more groups of oral diabetes medications. The dataset contains five classes of oral diabetes medication as shown in Table 1 [14]. Thus, we replaced 24 features for diabetes medications with two features, namely "Number of medications classes" and "Insulin". A number of medications classes feature indicates how many groups of oral medications are used by the patient.

Table 1: Classes of Oral Diabetes Medications [14]

Oral Medications Class	Medication Name
Thiazolidinedione	Pioglitazone, Rosiglitazone, and Troglitazone
Alpha-glucosidase Inhibitor	Acarbose and Miglitol
Sulfonylureas	Glimepiride, Glyburide, Chlorpropamide, Tolbutamide, Glipizide, Acetohexamide, and Tolazamide
Meglitinide	Nateglinide and Repaglinide
Biguanide	Metformin
Combination Medicines	Glyburide-metformin, Glipizide-metformin, Glimepiride-pioglitazone, Metformin-rosiglitazone, Metformin-pioglitazone

Upon preprocessing and eliminating the features that were not related, the dataset is ended up with 84749 instances and 16 features.

5. THE EFFECT OF THE AGENT

In this section, the research studies the effect of using the agent for big data preprocessing on two factors: the memory capacity and data transfer time. As mentioned earlier, research experiments are conducted on diabetes database collected from 130 US hospitals for years 1999-2008. It includes 74,036,643 instances and 117 features. It consists of a wide variety of heterogeneous data including structured (such as patient demographic data and lab results), semi-structured, and unstructured (such as paper prescriptions, physician notes, radiography films, MRI (magnetic resonance imaging), CT (computed tomography) and other images).

A. Agent Effect on Memory Capacity

In the healthcare field, due to the diversity of the medical records, the heterogeneity of healthcare, the capacity of healthcare data is constantly growing. This would lead to a dramatic increase in the cost of the data storage. Researchers believe that using agents for preprocessing the data can overcome this problem. To show up this, computing the size of the dataset before and after preprocessing using an agent via (1):

The data size = number of instances * number of features * average cell size (1)

Where the size of the cell on average is 1 KB. This yields 8.6623 terabytes (TB). After preprocessing phase using the agent, the dataset has 101766 instances and 50 features. By using equation 1, the dataset size is 5.0883 GB.

B. Agent Effect on Data Transfer Time

This part is testing how the agent affects the data transfer time over a network. As mentioned before, the dataset size before using the agent is 8.6623 TB. The dataset was collected from 130 US hospitals, Thus, the time needed to transfer each dataset over a 507 kilobits/s (Kbps) communication link is 12 days: 4 hours: 3 minutes: 29 seconds.

After using the agent, the dataset size is 5.0883 GB. For each dataset, the actual time needed for data transfer over a 507 Kbps communication link is 10 minutes: 17 seconds.

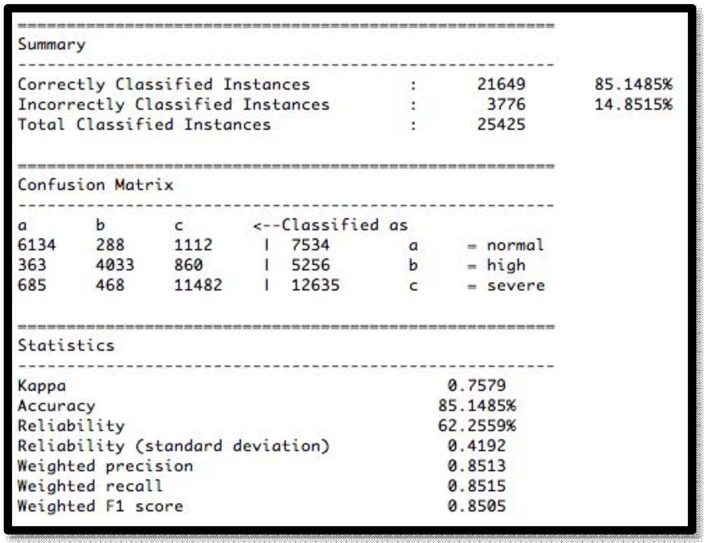
As noted above, the agent can significantly reduce the time taken for transferring data over a network.

6. RESULTS AND ANALYSIS

In this section, classification results for the dataset are represented. Then, the prediction results based on each single attribute in the dataset are presented.

A. Big Data Classification Results

The results obtained by applying the proposed model to the diabetes dataset described earlier are explained in this section. These results are represented in the form of a confusion matrix as shown in Figure 3.



```
=======================================================
Summary
-------------------------------------------------------
Correctly Classified Instances         :     21649        85.1485%
Incorrectly Classified Instances       :      3776        14.8515%
Total Classified Instances             :     25425

=======================================================
Confusion Matrix
-------------------------------------------------------
a       b       c       <--Classified as
6134    288     1112     |  7534        a       = normal
363     4033    860      |  5256        b       = high
685     468     11482    |  12635       c       = severe

=======================================================
Statistics
-------------------------------------------------------
Kappa                                          0.7579
Accuracy                                      85.1485%
Reliability                                   62.2559%
Reliability (standard deviation)               0.4192
Weighted precision                             0.8513
Weighted recall                                0.8515
Weighted F1 score                              0.8505
```

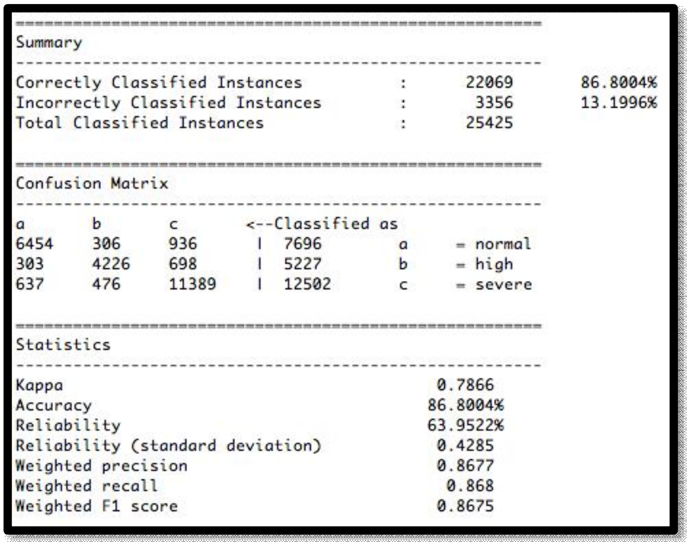
Figure 3: Classification results of proposed Model

As the Figure shows, the test data contained 25425 instances, where 21649 instances of them (i.e. 85.1485%) are correctly classified, whereas 3776 instances (i.e. 14.8515%) are incorrectly classified. In other words, out of 7534 normal cases, 6134 are correctly classified as normal, 288 are incorrectly classified as high, and 1112 are incorrectly classified as severe. Out of 5256 high cases, 4033 are correctly classified as high, 363 are incorrectly classified as normal, and 860 are incorrectly classified as severe. Out of 12635 severe cases, 11482 are correctly classified as severe, 685 are incorrectly classified as normal, and 468 are incorrectly classified as high.

Overall, the model correctly classifies 81.42% of normal cases, 76.73% of high cases, and 90.87% of severe cases, which is considered the most important class.

Other experiments are carried out and their effect on classification performance is examined. In this research, we examine some attributes to determine if there are any unbalance distribution values. The attribute “race” is removed from the dataset since it has unbalanced distribution. The “race” attribute has about three-quarter of the total number of instances represented by one class, and about 95% of the total number of instances represented by two classes.

There is another important experiment, which the researchers believe it may improve the classification results. The ten classes for the “age” attribute are consolidated into only three classes that are: young (0-40), middle-aged (40-60), and old (60 and over). We suggest this new grouping because the three new classes are better representing the age stages. The classification results are shown in Figure 4.



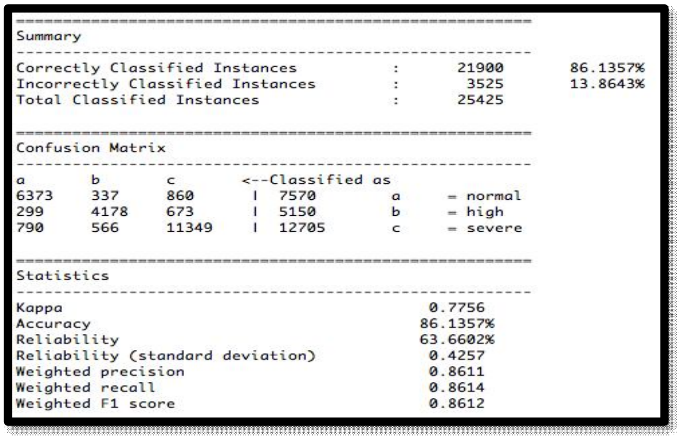
```
=======================================================
Summary
-------------------------------------------------------
Correctly Classified Instances         :     22069        86.8004%
Incorrectly Classified Instances       :      3356        13.1996%
Total Classified Instances             :     25425

=======================================================
Confusion Matrix
-------------------------------------------------------
a       b       c       <--Classified as
6454    306     936      |  7696        a       = normal
303     4226    698      |  5227        b       = high
637     476     11389    |  12502       c       = severe

=======================================================
Statistics
-------------------------------------------------------
Kappa                                          0.7866
Accuracy                                      86.8004%
Reliability                                   63.9522%
Reliability (standard deviation)               0.4285
Weighted precision                             0.8677
Weighted recall                                0.868
Weighted F1 score                              0.8675
```

Figure 4: Results after the regrouping of age

From Figure 4, it is obvious that there is an improvement in kappa, accuracy, reliability, standard deviation, weighted precision, weighted recall, and weighted F1 score. For example, the model accuracy was 85.1485% and became 86.8004% with the new age grouping.

B. Prediction Based on Single Attribute

This section studies the effect of each single attribute in the dataset on the classification process. Figure 5 presents the results of classification based on the “gender” attribute.



```
=======================================================
Summary
-------------------------------------------------------
Correctly Classified Instances         :     21900        86.1357%
Incorrectly Classified Instances       :      3525        13.8643%
Total Classified Instances             :     25425

=======================================================
Confusion Matrix
-------------------------------------------------------
a       b       c       <--Classified as
6373    337     860      |  7570        a       = normal
299     4178    673      |  5150        b       = high
790     566     11349    |  12705       c       = severe

=======================================================
Statistics
-------------------------------------------------------
Kappa                                          0.7756
Accuracy                                      86.1357%
Reliability                                   63.6602%
Reliability (standard deviation)               0.4257
Weighted precision                             0.8611
Weighted recall                                0.8614
Weighted F1 score                              0.8612
```

Figure 5: Classification results based on gender

Figure 6 summarizes the classification results of each individual attribute.

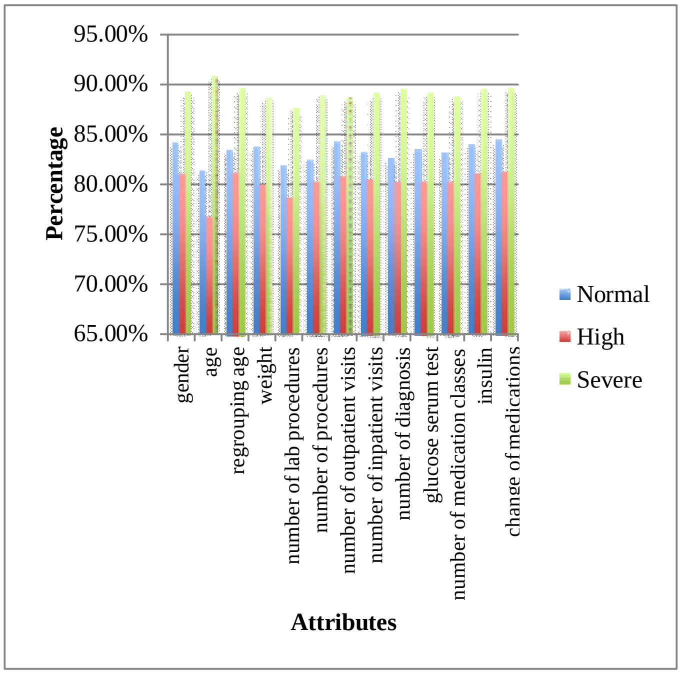


Figure 6: Classification results of each attribute

7. CONCLUSION AND FUTURE WORK

It is the era of big data, where more information is being captured in ever-finer detail from more sources than ever before. The main challenge for the big data applications is to explore such data and extract useful knowledge for future actions. Therefore, designing and developing applications to handle big data has attracted the researchers of this paper.

One of the most important representations of big data is healthcare applications. Diabetes is the most widely recognized endocrine ailment found in all populace and age gatherings. On the off chance that diabetes isn't dealt with appropriately and on time, it can prompt horrendous long haul complexities. This makes diabetes one of the primary needs in medicinal science examine.

This research developed and implemented an agent-based big data classification model and studied its effect in healthcare field.

The model efficiently classified diabetic patients into three different levels based on severity with 85.8794% average accuracy.

The research has a lot of open research areas that may lead to good research points in the future: Using the proposed model to predict the severity of other diseases and trying to perform the mining process remotely at the data sides rather than locally and comparing the results. Also, building the model with other classification algorithms may yield different

REFERENCES

[1] Rahman A. and Tasnim S., “Ensemble Classifiers and Their Applications: A Review,” *International Journal of Computer Trends and Technology (IJCTT)*, vol. 10, no. 1, pp. 31-35, Apr. 2014.
https://doi.org/10.14445/22312803/IJCTT-V10P107

[2] Breiman L., “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5- 32, Jan. 2001.
https://doi.org/10.1023/A:1010933404324

[3] Nithya J. and Geetha R., “Agent-Based Data Mining In Mobile Commerce: An Overview,” *International Journal of Computer Science Engineering and Technology (IJCSET)*, vol. 2, pp. 1065-1068, Apr. 2012.

[4] Song D., Lau R., Bruza P., Wong K.-F., and Chen D.-Y., “An Intelligent Information Agent for Document Title Classification and Filtering in Document-intensive Domains,” *Decision Support Systems*, vol. 44, no. 1, pp. 251-265, Nov. 2007.
https://doi.org/10.1016/j.dss.2007.04.001

[5] Kalaivani P. and Shunmuganathan K.L., “An Agent Based Framework for Sentiment Classification of Online Reviews Using Ontology,” *Journal of Computer Science*, vol. 10, no. 5, pp. 809-820, 2014.
https://doi.org/10.3844/jcssp.2014.809.820

[6] Pinzon C., Paz J., Rodriguez S., Bajo J., and Corchado J., “A Hybrid Agent-based Classification Mechanism to Detect Denial of Service Attacks,” *Journal of Physical Agents*, vol. 3, no. 3, pp. 11-18, Sept. 2009.

[7] Zolfaghar K., Meadem N., Teredesai A., Roy S., Chin S.-C., and Muckian B., “Big Data Solutions for Predicting Risk-of-Readmission for Congestive Heart Failure Patients,” in *Proc. of The 2013 IEEE International Conference on Big Data*, 2013, pp. 64-71.
https://doi.org/10.1109/BigData.2013.6691760

[8] Ko K., El-Ghazawi T., Kim D., and Morizono H., “Predicting The Severity of Motor Neuron Disease Progression Using Electronic Health Record Data with a Cloud Computing Big Data Approach,” in *Proc. of the 2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, 2014, pp. 1-6.
https://doi.org/10.1109/CIBCB.2014.6845506

[9] (2014) PRO-ACT. [Online]. Available: https://nctu.partners .org/ ProACT/

[10] Kiruba H.R. and Arasu G.T., “An Intelligent-Agent Based Framework for Liver Disorder Diagnosis Using Artificial Intelligence Techniques,” *Journal of Theoretical and Applied Information Technology*, vol. 69, no. 1, pp. 91-100, Nov. 2014.

[11] Kulothungan K., Ganapathy S., Yogesh P., and Kannan A., “An Agent Based Intrusion Detection System for Wireless Sensor Networks Using Multilevel Classification,” *International Journal of Modern Engineering Research (IJMER)*, vol. 2, no. 1, pp. 55-59, 2012.

[12] Strack B., DeShazo J., Gennings C., Olmo J., Ventura S., Cios K., and Clore J., “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical

Database Patient Records," *BioMed Research International*, vol. 2014, pp. 1-11, Apr. 2014. https://doi.org/10.1155/2014/781670

[13] (2014) Diabetes 130-US hospitals for years 1999-2008 Data Set. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

[14] Harper W., Clement M., Goldenberg R., Hanna A., Main A., Retnakaran R., Sherifali D., Woo V., and Yale J.-F., "Pharmacologic Management of Type 2 Diabetes," *Canadian Journal of Diabetes*, vol. 37, no. 1, pp. S61-S68, Apr. 2013 https://doi.org/10.1016/j.jcjd.2013.01.021