DATA MINING USING BUSINESS INTELLIGENCE AND SQL

Ayanesh Chowdhury[1], Rachakonda Aditya Vardhan[2], KASI KOUSHIK[3], Sheethal Dongari[4], Hrushikesh Sriramaneni[5]

Abstract: Most packages structures keep, get proper of entry to and control their facts the usage of relational information bases. This studies goals to construct a Classifier tool that analyzes and mines information the use of Standard Query Language (SQL). Specifically, this paper discusses a Music Genre Classifier machine that makes use of a relational database to accept tuples of audio abilities as input data and then makes use of a model that modified into constructed the usage of a records mining device but emerge as parsed and transformed to SQL statements to are expecting the beauty labels of musical compositions. In building the Music Genre Classifier device, jAudio a Digital Signal Processing device turn out to be used to preprocess the enter records through the extraction of audio capabilities of musical compositions (songs); WEKA changed into used to discover several information mining algorithms and to construct the prediction model; MS Access emerge as used to simply accept inputs in relational format and to execute the prediction model in SQL. Classification, clustering, and association rule mining algorithms in WEKA had been studied, explored, in comparison and then the most suitable approach turned into determined on to increase the system. Particularly, best algorithms that generated desire wood and policies as fashions were considered for the cause that the ones sorts of output can be without problem parsed and then transformed to SQL statements. This paper additionally discusses how desire trees and policies generated from WEKA are parsed and converted to SQL statements. For the comparative evaluation of the numerous algorithms that have been considered, experiments to test and measure their predictive accuracy had been performed. For the classifiers, J48 acquired the first-class predictive accuracy; for the Clusterers, Simple K-Means with J48 produced the highest predictive accuracy; and for Association, Predictive Apriori has the best accuracy fee. Overall, J48 stood out to be the high-quality algorithm for prediction of musical genre.

1.INTRODUCTION

Data mining that is a confluence of many disciplines can be described in severa methods. According to [Rajaraman et al., 2011], the maximum normally everyday definition of "statistics mining" is the discovery of "fashions" for data. As a famous era, statistics mining may be completed to any type of facts (e.G., data streams, ordered/collection records, graph or networked statistics, spatial records, text statistics, multimedia data, and the WWW) as long as the information are extensive for a aim utility [Han, et al., 2012].

Most packages structures use traditional databases to preserve, access and manipulate data. To widen the appeal of facts mining to the developer and customer groups, statistics mining software systems

Should be handy to apply and be without problems deployable in real-international surroundings. Central to carrying out this objective is the combination of information mining with conventional database structures [Chaudhuri, et al., 2001].

This observe makes a speciality of the mining of multimedia statistics, audio particularly. We have built a Music Genre Classifier gadget that predicts or classifies the fashion of an unclassified musical piece on top of a relational database. Basically, the Music Genre Classifier machine employs jAudio to extract the relevant functions of an unclassified musical piece. The device accepts the ones capabilities and stores them in MS Access, a relational database tool. A SQL question announcement, which at the begin turned into a prediction version generated by manner of WEKA, but became converted and coded in SQL, is then finished to are looking ahead to the fashion of the musical piece. In constructing the prediction models, 3 types of information mining strategies had been examined, in particular, kind, clustering and association rule mining. WEKA, a statistics mining workbench, was applied for this purpose. It has several integrated machine studying algorithms which can generate prediction models from a schooling fact set. However, research of the algorithms become confined to handiest people who produce policies and choice wood as prediction fashions. Rules and choice bushes can be without difficulty parsed and translated to SQL query statements.

Our study is prepared and furnished as follows. Section discusses facts preprocessing, where the audio documents are organized for statistics mining. In segment 3, the various strategies that have been explored to construct the statistics fashions for prediction are mentioned. Also, the consequences of the experiments executed are supplied. Lastly, in phase 4, conclusions of the have a look at are drawn and suggestions for destiny research are provided.

2.DATA-PREPROCESSING

Data preprocessing includes the transformation of uncooked records into an understandable format. It prepares raw information for facts mining [Technopedia, n. D.]. For track, records including assault, duration, quantity, tempo and device

form of every single be aware are to be had. Statistical measures along with pace and advise key for every tune item can without difficulty be extracted [Kotsiantis et al., 2004] and in this observe jAudio changed into used.

The jAudio [Sourceforge, n. D.] is a Digital Signal Processing gadget that lets in clients to extract audio competencies or homes which consist of beat factors, statistical summaries, and many others. It has a GUI, an API for embedding jAudio in programs and a command line interface for facilitating scripting. It has functionalities that allow customers to set fashionable parameters consisting of window size, window overlap, down sampling and amplitude normalization. It can also carry out audio synthesis, document audio and switch MIDI documents to audio. It has the functionality to display audio signals in both frequency and time domain names. It can parse MP3, WAV, AIFF, AIFC, AU and SND documents. It allows function values to be created in either ACE XML or WEKA ARFF documents [McKay, C., 2010].

In this look at, nice those capabilities deemed applicable to track style type have been extracted from jAudio. These functions encompass Spectral Centroid, it is a diploma of the "centre of mass" of the power spectrum; Spectral roll off factor, which is a degree of the quantity of the right-skewedness of the energy spectrum; Spectral flux, which is a amazing degree of the quantity of spectral exchange of a signal; Compactness, which is a splendid degree of ways crucial a feature everyday beats play in a piece of track; Spectral variability, that is a diploma of approaches various the importance spectrum of a sign is; Root mean square (RMS), that is a great degree of the electricity of a sign; Fraction of low strength home windows, which is a top notch degree of the way quite a few a sign is quiet relative to the relaxation of a signal; Zero crossings, which is a wonderful diploma of the pitch further to the noisiness of a sign; Strongest beat, this is strongest beat in a signal and it's far discovered by way of manner of locating the very quality bin within the beat histogram; Beat sum, that's a terrific diploma of ways important a function ordinary beats play in a piece of tune; Strength of strongest beat, that could be a diploma of how sturdy the maximum powerful beat is in assessment to different possible beats; MFCC, it's a degree of the coefficients that make up the fast time period energy spectrum of sound; LPC, which calculates linear predictive coefficients of a sign; and Method of moments, that's a much like Area Method of Moments function, however does now not have the big offset [Sourceforge, n. D.]. Two extra abilties or attributes for each tune were brought: Class, it really is the genre of the song and the ID, which uniquely identifies a music (for reference functions).

In this take a look at, best five musical genres have been considered, specially, classical, u.S.A. Of the usa, jazz, reggae and rock. These had been decided on due to the fact they may be adjudged to have the maximum top notch beats and Features. A preferred of 622 songs have been gathered and preprocessed (characteristic extraction); 512 of which have been used as schooling records and the rest (a hundred and ten) had been used as take a look at statistics. The schooling facts set consists of one hundred classical, 100 united states, 100 jazz, one hundred reggae, one hundred rock, 6 classical-rock, 6 u . S . A .-rock. The test records set consists of 20 classical, 20 usa of america, 20 jazz, 20 reggae,

20 rock, 5 classical-rock, 5 usa-rock. The extracted features for every training and take a look at facts units have been created as ARFF documents.

Three. BUILDING DATA MODELS

Essentially, building records fashions for prediction involves the application of statistics mining strategies that generate significant patterns. In this studies, the statistics mining strategies which have been taken into consideration and studied include magnificence, clustering and association rule mining. To aid us within the research of these many exceptional techniques, we used WEKA, that is a information mining workbench that has advanced immensely in its information mining abilties. Incorporated in WEKA is an high-quality range of tool reading algorithms and associated techniques. It now consists of many new filters, device mastering algorithms, and attributes choice algorithms, and many new additives collectively with converters for one-of-a-kind file formats and parameter optimization algorithms. [Witten, et al, 2011] As defined in [Abernethy, 2010], WEKA is the fabricated from the University of Waikato (New Zealand) and changed into first applied in 1997. It uses the GNU General Public License (GPL).

The software program application is written inside the Java™ language. It includes a GUI for interacting with records documents and producing visible effects (assume tables and curves). It moreover has a popular API, which allows developers to embed WEKA in applications. In terms of beneficial components, WEKA has three graphical consumer interfaces, especially, the explorer, experimenter, and knowledge waft and a command line interface. The explorer GUI has six panels which constitute a facts mining mission – preprocess, classify, cluster, partner, pick attributes and visualize [The University of Waikato, 2008; Witten, et al., 2011]. In this have a look at, maximum of the paintings finished come to be undertaken within the explorer interface, and some thru WEKA's command line interface. As cited in phase 1, we restricted the scope of our research of statistics mining algorithms. We first-class considered those algorithms that generate policies and choices timber as fashions for prediction.

Primarily, we did so due to the fact rules and choice bushes can be without troubles parsed and translated to SQL query statements.

For classification and association rule mining, the technology of models (in the form of policies and preference trees) and their conversion to SQL query statements is pretty easy. However, for clustering, additional steps had been undertaken. Since clustering algorithms in WEKA do not generate hints or choice trees, we as a substitute used cluster analysis as a preprocessing method in which each pattern data in the training set is grouped right right into a cluster. We then done J48 to the clustered education facts to produce the selection tree.

The succeeding subsections discuss the algorithms that had been investigated and gift the outcomes of the experiments finished.

Three.1Classification

According to [Han, et al., 2012], magnificence is the manner of locating a model (or function) that describes and distinguishes records lessons or ideas. The model is generated based totally on the assessment of a hard and rapid of schooling information (i.E., information devices for which the elegance labels are acknowledged) and is used to predict the elegance label of unclassified gadgets.

The category algorithms decided on are J48, BFTree and RandomTree. For each algorithm, one model in step with style is built. In preferred, there have been forty five fashions generated.

The J48 Decision tree classifier is based on C4.Five, an set of guidelines that changed into evolved via J. Ross Quinlan. In J48 set of regulations as described in [Padhye, n. D.], if you want to classify a trendy object, a selection tree primarily based at the attribute values of the available training records is created first. Whenever a fixed of objects (education set) is encountered, an attribute that discriminates the various times most in fact is identified. This function is referred to as records advantage. It is used to determine the satisfactory manner to categorise the statistics. Among the possible values of this option, if there may be any cost for which there's no ambiguity, the branch is terminated after which goal price that become obtained is assigned to it. For the other cases, another characteristic that gives the very exceptional statistics gain is searched. The new launch keeps till a clean desire of what mixture of attributes gives a selected target charge is obtained, or while all attributes has been exhausted. In the occasion that attributes had been exhausted, or unambiguous end result from the available records can not be acquired, a goal fee that the majority of the devices underneath this department very own is assigned to this department.

On the opposite hand, the BFTree algorithm builds a high-quality-first desire tree classifier. It makes use of binary cut up for each nominal and numeric attributes. For lacking values, the technique of 'fractional' times is used [Theofilis, 2013]. Meanwhile, the Random Tree set of rules constructs a tree that considers K randomly chosen attributes at each node. It does not carry out pruning [Theofilis, 2013].

Table 1. Comparison of Classifiers: Percentage Table

	Kappa	TP Rate	FP Rate
BFTree	0.61	0.88	0.30
Random Tree	0.60	0.87	0.29
J48	0.63	0.88	0.30

Table 2. Comparison of Classifiers: Kappa statistics, TP rate

	J48	BF Tree	Random Tree
Classical	90.91	90	90.91
Country	85.45	88.18	80.91
Jazz	90	85.45	91.82
Reggae	82.87	84.55	84.55
Rock	88.18	90	87.27

The Kappa statistic or Kappa coefficient is used to diploma the agreement among expected and located categorizations of a dataset, on the same time as correcting for an agreement that takes region by means of the usage of threat. A kappa value of 1 shows best settlement whilst a kappa cost of 0 indicates settlement that is equal to danger [Witten et al., 2011; the University of Waikato, n. D.]. Based on the effects shown, the kappa rate of J48 is maximum. It also has the very exceptional TP Rate, the lowest FP charge and maximum percent of predictive accuracy. Thus, some of the 3 kind algorithms that were investigated, J48 stood out to be the fine.

3. Association Rules

In [Rouse, M. 2011] affiliation policies are described as though/then statements that assist reveal relationships among seemingly unrelated facts in a facts set. An affiliation rule includes elements, an antecedent (if) and a consequent (then). An antecedent is an item located within the information at the same time as a consequent is an object that is observed in mixture with the antecedent. Association rules are generated through studying information for frequent if/then styles and using the standards help and confidence to pick out the most critical relationships. Support is a degree of ways regularly the gadgets appear in the database. On the alternative hand, self belief shows the wide type of instances the if/then statements were found to be proper. In information mining, association rules can at instances be used for prediction [Deogun et al., 2005].In this test the association rule mining algorithms that were selected are Apriori, Filtered Associator, and Predictive Apriori. As defined in [Han, et al., 2012], the Apriori is a seminal set of policies that "employs an iterative method referred to as a degree-practical are searching for, wherein k-itemsets are used to find out (ok + 1)-itemsets. First, the set of common 1-itemsets is found via scanning the database to accumulate the depend for every item, and gathering those gadgets that satisfy minimum aid. The ensuing set is denoted with the aid of manner of L1. Next, L1 is used to find out L2, the set of common 2-itemsets, this is used to discover L3, and so on, till no more commonplace ok- itemsets may be discovered. The finding of each Lkrequires one complete test of the database. To improve the performance of the quantity-sensible era of common itemsets, an critical belongings referred to as the Apriori belongings is used to lessen the quest location." The Filtered Associator is an algorithm for "strolling an arbitrary associator on statistics that has been exceeded via an arbitrary filter out. Like the associator, the shape of the filter out is based totally completely at the training dataset and check instances can be processed by way of the use of the filter with out converting their form" [Knime n. D.]. On the alternative hand, PredictiveApriori algorithm "searches with an increasing useful resource threshold for the great 'n' guidelines concerning a help- based totally corrected self belief fee "[Knime (2), n. D.]. In WEKA, the affiliation policies aren't used for prediction. To be capable of use them for prediction we converted the association tips produced with the aid of the 3 algorithms to SQL query statements. These statements had been then completed to predict the style of each musical piece within the take a look at data set.

Table 3 shows a assessment of the proportion predictive accuracy.

Association rule algorithms the use of SQL query statements. Based on the consequences proven in Table 3, Predictive Apriori has the best percent of predictive accuracy.

A assessment of Table 1 and Table 3 suggests that classification registered better consequences than affiliation. It should however be referred to that the lower predictive accuracy of affiliation rules is in particular due to the unfinished technology of the rules due to the lack of computing assets.

	Apriori	Filtered Associator	Predictive Apriori
Classical	60	60	80
Country	44	44	44
Jazz	45	45	45
Reggae	60	60	60
Rock	70	80	83.33

Table 3. Comparison of Association Algorithms: Percentage Prediction Accuracy Using SQL

3.1 Clustering

Clustering diverges from classification and regression. While each classification and regression examine class classified (education) information units, clustering alternatively analyzes information objects without consulting magnificence labels. In many instances wherein elegance classified statistics may additionally in reality no longer exist at the beginning, clustering may be utilized to produce class labels for a set of statistics [Han, et al., 2012].

In clustering, "the objects are grouped based totally at the precept of maximizing the intraclass similarity and minimizing the interclass similarity. That is, clusters of gadgets are shaped so that objects within a cluster have high similarity in contrast to each other, however are as a substitute varied to gadgets in other clusters. Each cluster so formed may be viewed as a category of items, from which policies may be derived" [Han, et al., 2012].

Since clustering algorithms in WEKA, do now not generate regulations or decision timber, in our observe, we used clustering as a way to preprocess the training statistics. Based on the outcome of the cluster analysis that changed into

undertaken, we relabeled the elegance characteristic of each example in the training information set. The connected style to every musical piece changed into dropped in favor of the cluster group.

The clustering algorithms used are EM (expectation-maximization), Make Density Based Clusterer and Simple K-Means. As described in [Wikipedia, n. D.], "the EM set of rules is used to find the most likelihood parameters of a statistical version in cases wherein the equations can't be solved at once." [Wikepedia-1, n. D.]. On the opposite hand, according to [Wikepedia-2, n. D.], "in density-based totally clustering, clusters are defined as areas of better density than the remainder of the facts set. Objects in these sparse regions - which can be required to split clusters - are generally considered to be noise and border factors." Meanwhile, "K-method clustering aims to partition n observations into okay clusters in which each statement belongs to the cluster with the nearest imply, serving as a prototype of the cluster. This effects in a partitioning of the data space into Voronoi cells" [Wikepedia-3, n. D.].

After performing cluster evaluation on the schooling statistics, simplest 2 clusters (this is, a song is both categorized as classical or non-classical) were formed. The clustering algorithms couldn't virtually distinguish the other four musical genres from each other. Consequently, we relabeled the elegance labels (or genre) of the education facts as either belonging to classical or non-classical. The J48 which has the highest percentage of predictive accuracy a few of the class algorithms become then implemented to generate the choice timber.

Table 4. Comparison of Cluster Analysis: Percentage Prediction Accuracy using J48

	EM	Simple K-Means	Make Density Based Clusterer
Classical	92.73	92.73	95.45
NonClassical	92.50	93.64	92.28

Table 4 shows a comparison of the share predictive accuracy of J48 the usage of the 3 preprocessed (thru clustering) records units. Based on the consequences proven, Simple K-Means with J48 produced the highest percent of predictive accuracy. However, it must be mentioned that Country, Jazz, Reggae and Rock have been all clustered as non- classical.

4. GENRE PREDICTION USING SQL

This segment shows a snippet of the algorithm that become used to parse and convert the choice tree to SQL Query statements (see Fig 1). Fig 2 suggests a sample of a decision tree produced with the aid of J48. Fig 3 shows a snippet of the SQL Query statement that turned into generated via the set of rules proven in Fig. 1

```
while (scanner.hasNext() && countGenre < total) {
    orCount = 0;
    attribute = "";
    condition = "";
    token = scanner.next();
    if (token.equals("|")) {
        orCount++;
        token = scanner.next();
        while (token.equals("|")) {
            orCount++;
            token = scanner.next();
        }
        if (temp.isEmpty() && prev != null) {
            temp = new ArrayList<Line>(prev.subList(0, orCount));
        }
    }
    attribute = token;
    token = scanner.next();
    while (!(token.equals("<=")) && !(token.equals(">")) && !(token.equals(">=
&& !(token.equals("<")) && !(token.equals("=")))) {
        attribute = attribute + "_" + token;
        token = scanner.next();
    }
    condition = token;
    token = scanner.next();
```

```
Spectral Flux Overall Standard Deviation0 <= 0.001399
|   Spectral Rolloff Point Overall Standard Deviation0 <= 0.1129
|   |   LPC Overall Average0 <= -0.958
|   |   |   Spectral Centroid Overall Standard Deviation0 <= 4.231: Classical (2.0)
|   |   |   Spectral Centroid Overall Standard Deviation0 > 4.231: notClassical (4.
|   |   LPC Overall Average0 > -0.958: Classical (92.0)
|   Spectral Rolloff Point Overall Standard Deviation0 > 0.1129: notClassical (6.0/1
Spectral Flux Overall Standard Deviation0 > 0.001399
|   Method of Moments Overall Average3 <= 182300
|   |   Method of Moments Overall Standard Deviation0 <= 0.1137: Classical (4.0)
|   |   Method of Moments Overall Standard Deviation0 > 0.1137
|   |   |   LPC Overall Average2 <= -0.4737: Classical (3.0)
|   |   |   LPC Overall Average2 > -0.4737
|   |   |   |   MFCC Overall Standard Deviation1 <= 7.117: notClassical (39.0/1.0)
|   |   |   |   MFCC Overall Standard Deviation1 > 7.117: Classical (3.0/1.0)
|   Method of Moments Overall Average3 > 182300: notClassical (359.0/1.0)
```

```
SELECT * FROM MusicData WHERE (LPC_Overall_Standard_Deviation5 >
0.1178 and Spectral_Flux_Overall_Average0 <= 0.00391 AND
MFCC_Overall_StandardDeviation9 <= 2.432 AND
Spectral_Rolloff_Point_Overall_Standard_Deviation0 <= 0.207 AND
LPC_Overall_Standard_Deviation1 <= 0.2466 AND MFCC_Overall_Average11
<= 0.6598 AND Compactness_Overall_Standard_Deviation0 > 191.5) OR
```

5. CONCLUSIONS AND RECOMMENDATIONS

In this examine, we were able to expose that SQL Query statements may be used for style prediction by means of changing decision tree and rule-primarily based models generated by using WEKA to SQL statements. We have been able to integrate of facts mining with traditional database systems.

Among the information mining strategies of class and association rule mining that had been investigated, this take a look at located that based on the effects of the experiments carried out, J48 category algorithm has the highest percentage of predictive accuracy.

It become also observed that most in all likelihood due to the small schooling data set used on this take a look at, clustering algorithms were simplest capable of discover 2 clusters of facts (that is, a musical piece is both classified as classical and non-classical). As a effect, clustering as a preprocessing method turned into not confirmed to be useful on this specific case.

The Association policies approach has registered the lowest percent of predictive accuracy and this changed into because of some barriers encountered for the duration of the observe. WEKA required a number of reminiscence potential whilst processing affiliation rule algorithms. Due to the dearth of extra powerful computing assets, the generation of affiliation rules became now not completed; handiest selected rules were used for prediction.

In the destiny, to provide more meaningful consequences, we propose the use of a bigger schooling facts set. In the examine performed by [McKay, 2004], extra than 35,000 songs were to be had for education facts.

6. REFERENCES

1) Abernethy, M. (2010). Data mining with WEKA, Part Introduction and regression. Developer Works. IBM. Retrieved January 29, 2012 from http://www.ibm.com/developerworks/library/os-weka1/

2) Chaudhuri, S. ; Fayyad, U. ; Bernhardt, J. (2001). Integrating data mining with SQL databases: OLE DB for data mining. Proceedings of the 17th International Conference on Data Engineering, 2001. pp. 379 – 387.

3) Deogun, J. and Jiang, L. (2005) Prediction Mining – An Approach to Mining Association Rules for Prediction. Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing Lecture Notes in Computer Science Volume 3642, 2005, pp 98-108

4) Han, J., Kamber, M. and Pei, J. (2012). Data Mining: Concepts and Techniques 3rd Edition, Morgan Kaufmann

5) Kotsiantis, S., Kanellopoulos, D. and Pintelas, P. (2004). Multimedia Mining. WSEAS Transactions on Systems 3 (10), 3263-3268

6) Knime. (n. d.). FilteredAssociator (3.6). Retrieved on January 29, 2014,from.http://www.knime.org/files/nodedetails/weka_associations_Filt eredAssociator.html

7) Knime(2). (n. d.). PredictiveApriori. Retrieved on January 29, 2014 from http://www.knime.org/files/nodedetails/weka_associations_Pre dictiveApriori.html

8) McKay, C. (2010). Automatic Music Classification with jMIR. PhD Dissertation. Department of Music Research Schulich School of Music McGill University, Montreal.

9) Padhye, A. (n. d.). Chapter 5. Classification Methods. Retrieved on January 29, 2014 from http://www.d.umn.edu/~padhy005/Chapter5.html

10) Rajaraman, A. & Ullman, J. (2010). Mining of Massive Datasets.(available for free on the web)

11) Rouse, M. (2011) Association Rules (in data mining). Search business analytics. Retrieved on January 29, 2014 from http://searchbusinessanalytics.techtarget.com/definition/associ ation-rules-in-data-mining

12) Sourceforge (n. d.), jAudio. Retrieved January 29, 2014 from http://jaudio.sourceforge.net/

13) Technopedia (n. d.). Data Preprocessing. Retrieved on January 31, 2014 from http://www.techopedia.com/definition/14650/data-preprocessing

14) The University of Waikato (2008). WEKA 3 – Data Mining with Open Source Machine Learning Software in Java. Source URL: http://www.cs.waikato.ac.nz/~ml/weka/The University of Waikato (n. d.). Primer. Retrieved on January 29, 2014 from http://weka.wikispaces.com/Primer

15) Theofilis, G. (2013). Weka Classifiers Summary. Retrieved on January 29, 2014 from http://www.academia.edu/5167325/Weka_Classifiers_Summary

16) Wikipedia-1. (n. d.) Expectation–maximization algorithm. Retrieved on January 31, 2014 from http://en.wikipedia.org/wiki/Expectation%E2%80%93maximiza tion_algorithm

17) Wikipedia-2. (n. d.). Cluster analysis. Retrieved on January 31, 2004 from http://en.wikipedia.org/wiki/Cluster_analysis

18) Wikipedia-3 (n. d.). k-means clustering. Retrieved on January 31, 2004 from http://en.wikipedia.org/wiki/K-means_clustering

19) Witten, I. and Frank, E., (2011). Data Mining Practical Machine Learning Tools and Techniques, 3rd Edition, Elseveir