International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Risk Distribution and Validation of Data in Passport Data Analysis Using Cluster Analysis

Sucheta Gulia[1], Dr. Rajan Vohra[2]

[1]M. Tech (C.E) Student, Department of Computer Science & Engineering,
P.D.M College of Engineering, Sector 3A, Sarai Aurangabad, Bhadurgarh, India

[2]Head of Department, Department of Computer Science & Engineering,
P.D.M College of Engineering, Sector 3A, Sarai Aurangabad, Bhadurgarh, India

Abstract: ***This paper presents a comprehensive statistical experiment to identify the sensitive office from the passport database. Data mining is the process of finding the meaningful patterns, correlations among dozens of fields that lie hidden within very large databases. The presented work focus on implementing different data mining approaches on passport database which is the primary data taken from passport offices. This paper combines two major approaches to provide profiling via clustering and validation of data using classification. According to defined approach, clustering is implemented to profile offices according to their risk identified. The design of experiments software named Weka Tool is used for making clusters using attributes place of issue and risk type dataset by using simple k means algorithm. With the simulation and analysis results, identify the sensitive centre which contribute to high risk score***

Keywords: Clustering(simple k-means), Profiling, Distribution, Validation, Weka tool

1. Introduction

A passport is a government-issued travel document that certifies the identity and nationality of its holder for the purpose of international travel. The elements of identity contained in all standardized passports include information about the holder, including name, date of birth, sex and place of birth.

A passport displays nationality, but not the place of residence of the passport holder. The passport holder is normally entitled to re-enter the country that issued the passport in accordance with the laws of that country, and in some instances of gaining a new citizenship, to enter that country for the first time. A passport does not necessarily grant the passport holder entry into any other country, nor to consular protection while abroad or other privileges, such as immunity from arrest or prosecution. Those rights and privileges, if and when applicable, arise from international treaties [1].

Clustering is the task of segmenting a diverse group into a number of similar subgroups or clusters. What distinguishes clustering from classification is that clustering does not rely on predefined classes. In clustering, there are no predefined classes. The records are grouped together on the basis of self-similarity. Clustering is often done as a prelude to some other form of data mining or modeling. For example, clustering might be the first step in a market segmentation effort, instead of trying to come up with a one-size-fits-all rule for determining what kind of promotion works best for each cluster.

Significance of the problem - This paper is used to find the sensitive centre which contributes to high risk score. This paper contains two parts, first part will help to find the sensitive office from where such data originates like if we use passport data from four passport offices then from this approach we are able to identify such centers from maximum fraudulent data entries originates. This is also used to check the risk but this is used for the passport offices to find the risk and the first is used on the database instances. In this k-means algorithm was used for clustering.

The second part is used to validate the data used in the database. This part gives the desired information about the incorrect instances of the database. The development of a commercial products based on this prototype will help us to identify fraudulent locations in real time.

2. Research Methodology

For solving the problem some research methodologies and algorithms are used for obtaining the result. Main steps under the research methodologies are as follow:-

1. Review literature or research papers – first of all literatures and research papers were reviewed for getting more information about the problem and knowing which type of work was done by others on this topic and by which method.
2. Identify tools – then tools required for solving the problem were identified and the best tool was selected from all.
3. Study database attributes and data structure – attributes and structure of the database was thoroughly studied for finding out useful attributes from the passport. For critical attributes used in the database first and last page of the passport was studied [1].
4. Organize filed visits to NIC (NATIONAL INFORMATICS CENTRE) New Delhi, INDIA, nodal agency for passport preparation. From there we get information about the flow of work for finding out sensitive applicants during passport generation.
5. Study the work flow in the work centre's like NIC(NATIONAL INFORMATICS CENTRE) and TCS(TATA CONSULTANCY SERVICE, SOFTWARE SOLUTION PROVIDER) New Delhi, INDIA.

6. Determine nature and definition of research problem and work flow of the problem for getting accurate and desired result.
7. Organize the database with useful attributes and populate it then perform data analysis using suitable tool e.g., WEKA in order to generate the result.

Classification consists of examining the features of a newly presented object and assigning to it a predefined class. The classification task is characterized by the well-defined classes, and a training set consisting of pre classified examples. The task is to build a model that can be applied to unclassified data in order to classify it. Examples of classification tasks include classification of credit applicants as low, medium or high risk.

J48 is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool. C4.5 is a program that creates a decision tree based on a set of labeled input data. This algorithm was developed by Ross Quinlan. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.

WEKA is a collection of machine learning algorithms for Data Mining tasks. It contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization. There was no preprocessing of the data. WEKA has four different modes to work in.

- Simple CLI; provides a simple command-line interface that allows direct execution of WEKA commands.
- Explorer; an environment for exploring data with WEKA.
- Experimenter; an environment for performing experiments and conduction of statistical tests between learning schemes.
- Knowledge Flow; presents a "data-flow" inspired interface to WEKA.

3. Research Background

Many popular classification methods like decision trees, neural network, and linear discriminates like Fisher's fall in this class. These differ a lot in what kind of model they produce and how they train such models but they all require the data to have a fixed set of attributes so that each data instance can be treated as a point in a multi-dimensional space. The training process partitions the space into regions for each class. When predicting the class label of an instance x, we use the defined region boundaries to find the region to which x belongs and predict the associated class. A number of methods have been applied for embedding sequences in a fixed dimensional space in the context of various applications. The simplest of these ignore the order of attributes and aggregate the elements over the sequence. For example, in text classification tasks a document that is logically a sequence of words is commonly cast as a vector where each word is a dimension and its coordinate value is the aggregated count or the TF-IDF score of the word in the document [4].

HMMs have been extensively used for modeling various kinds of sequence data. HMMs are popularly used for word recognition in speech processing [6]. [5] Report much higher classification accuracy with HMMs when used for detecting intrusions compared to previous k-grams approach. A lot of work has been done on building specialized hidden Markov models for capturing the distribution of protein sequences within a family [7].

Distance based clustering method is the most popular clustering method and includes the famous K-means and K-method clustering algorithms and the various hierarchical algorithms [8]. The primary requirement for these algorithms is to be able to design a similarity measure over a pair of sequences.

Data mining techniques like clustering and associations can be used to find meaningful patterns for future predictions [9, 10].Most methodology driven studies used mathematical methodologies; e.g. statistics, neural net, generic algorithm (GA) and Fuzzy set to identify the optimized segmented homogenous group [11-18].

Clustering has many applications, including part family formation for group technology, image segmentation, information retrieval, web pages grouping, market segmentation, and scientific and engineering analysis [19].

Many clustering methods have been proposed and they can be broadly classified into four categories [20-25].

4. Conceptual Framework

The research objectives of the paper are achieved by the following steps:

1) Collect information about passport holders and applicants then mine this database for getting necessary information only. By using different checks compute the score for each applicant. Then identify the risk and classify the applicants with different types of risk [1].
2) Then with the help of computed risk score and place of issue the distribution of the sensitive places takes place.
3) Then after distribution we can validate the data for assuring that the results obtained are correct.

4.1 Score Calculation and Risk Classification

Total score was calculated by summing the index check score, prior approval check score and police verification score. Then we get total score for every entry in the passport database. Then total score is computed simply getting sum of index check score, PAC score and PV score. Then according to score we will simply classify the score into low, medium and high risk [1].

4.2 Clustering and Distribution of Risk Score

In this clustering was done on the attribute place of issue and risk type. In this we assume the issuing places as place a, b, c, d and e. Through this technique different clusters are formed according to the risk type and issuing place like low risk on place a, b, c, d, and e, medium risk on place a, b, c, d, and e and high risk on place a, b, c, d, and e separately. Then using clustering distribution was done which gives information about the sensitive place from where the risk originates.

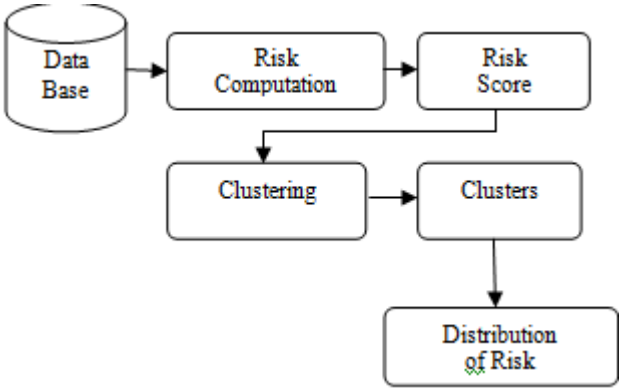


Figure 1: Clustering and Distribution

4.3 Validation of Risk Score

Risk score was validated using test data from the database using attribute place of issue and type of risk using the J48 technique. Validation gives the correctly placed instances and incorrectly placed instances in the database. Some basic steps are:-

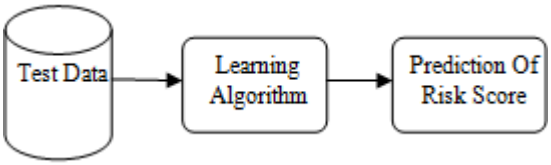


Figure 2: Validation of Risk Score

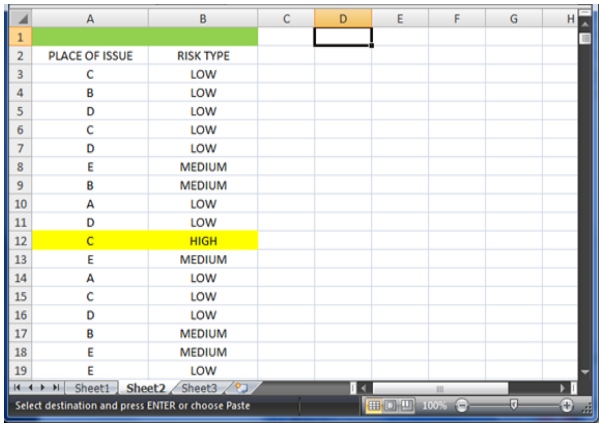
- Frame the test data.
- Trained the system for validation
- Use the suitable algorithm for the validation
- Anomaly Detection

5. Result and Discussion

The data mining method used to build the model is Clustering (K-mean clustering). The data analysis is processed using WEKAdata mining tool for exploratory data analysis.

Preparing the database [28] - A database of 478 records/entries is collected from nodal passport office. Primary data collected by random sampling, is used to solve the problem. The database consist of multiple attributes like, passport number, file number, social security number which is unique for every person, applicant name, father name, date of birth of the applicant, address, place of issue and different types of scores present in the databases of [1].

For obtaining the result for distribution of risk, data from the passport database was considered and two attributes were chosen from the database that are- Place of issue of the database and Risk type for the entry is fed to the Weka tool and as a result tool generate three clusters for each risk type corresponding to each passport issue. This was done using clustering method with K-means algorithm. After clustering different clusters were formed and give information about in which place the risk was low and in which the risk was high. This database was created from the passport database [28].For obtaining results for the second problem, the following database was created:



	A	B	C	D	E	F	G	H
1								
2	PLACE OF ISSUE	RISK TYPE						
3	C	LOW						
4	B	LOW						
5	D	LOW						
6	C	LOW						
7	D	LOW						
8	E	MEDIUM						
9	B	MEDIUM						
10	A	LOW						
11	D	LOW						
12	C	HIGH						
13	E	MEDIUM						
14	A	LOW						
15	C	LOW						
16	D	LOW						
17	B	MEDIUM						
18	E	MEDIUM						
19	E	LOW						

Sheet1 Sheet2 Sheet3
Select destination and press ENTER or choose Paste

Figure 3: Database used for Clustering showing Place of issue and Type of Risk

Figure 3 shows two attributes from passport database [1] i.e. place of issue and risk type. This dataset was formed after computation of score and identification of risk. From this dataset the entities with low risk, medium risk and high risks are separated. This separation was done using clustering technique.

Figure 4 shows the result obtained after clustering using the dataset shown in figure 3. Through clustering method we can identified the place from where maximum number of fraud entries or sensitive data originates.

5.1.1 Cluster 0 –
Cluster of applicants having High Risk.
Total Entries = 52 (11%)

5.1.2 Cluster 1 –
Cluster of applicants having Low Risk.
Total Entries = 276 (58%)

5.1.3 Cluster 2 –
Cluster of applicants having Medium Risk. Total Entries = 150 (31%)

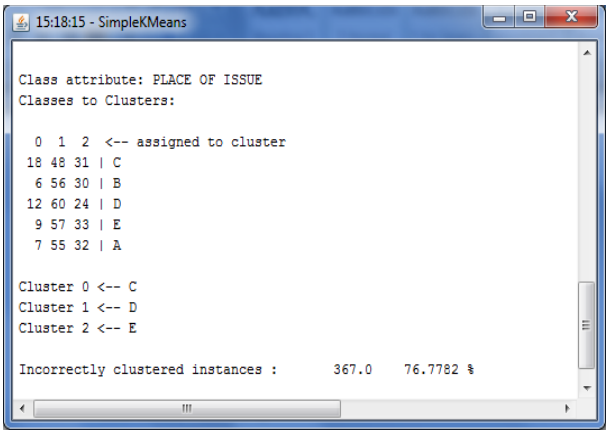


Figure 4: Classes to Cluster assignment (Result)

Cluster number 0 shows the high risk type applicants which are maximum in the passport office C. Cluster number 1 shows the Low risk type applicants and cluster 2 shows medium risk type applicants. The maximum number of entries for low risk is in passport office D and for medium risk are in passport office E. Figure 5 shows the graph

visualization of clusters which shows the entries of all the five passport offices according to the risk type.
B = 6 entries from cluster 0, 56 entries from cluster 1 and 30 entries from cluster 2.
D = 12 entries from cluster 0, 60 entries from cluster 1 and 24 entries from cluster 2.
E = 9 entries from cluster 0, 57 entries from cluster 1 and 33 entries from cluster 2.
A = 7 entries from cluster 0, 55 entries from cluster 1 and 32 entries from cluster 2.
C = 18entries from cluster 0, 48 entries from cluster 1 and 31entries from cluster 2.

Therefore, we get information from distribution technique that issuing place C has highest High Risk instances that are 18 and issuing place D has highest Low Risk instances that are 60. Hence through clustering the risk is distributed on different centres of issue from where such data originates.

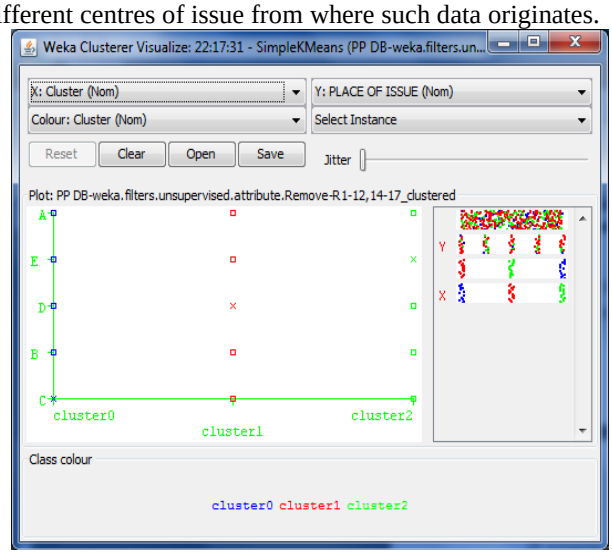


Figure 5: Graph visualizing of clusters

After distribution, prediction was done using J48 algorithm which can validate the data by giving information about the correctly placed instances and incorrectly placed instances.

This can be done by trained the system using test data and then J48 algorithm predict the instances as correct or incorrect. In test data, the database includes 10 instances with attributes place of issue, risk type and cluster. Validation of database can be done by using number of steps –

- Firstly the clusters are stored as ARFF data files.
- Load that file into the Weka tool. An instance number was created with the attributes, delete that instance number attribute.
- The ARFF data file gives another attribute which is instance number and the attributes given in the passport database.
- The ARFF file use two attributes, place of issue and Risk type and assign the cluster number to each entry of the database.
- Then test data was loaded into the tool by clicking on the supplied test data under the classify part of the Weka tool.
- Then choose the J48 algorithms from the tree section and click on start button. It can give the desired result in the form of correct and incorrect instances.

From this we can validate our data as the data in the data base is correctly placed or not. Figure 6 gives the confusion matrix during validation of data. This confusion matrix gives the output based on the test data which includes 10 entries. By using J48 algorithm matrix gives result in the form of correctly and incorrectly placed instances. In this 7 instances are correctly placed and 4 are placed at incorrect position.

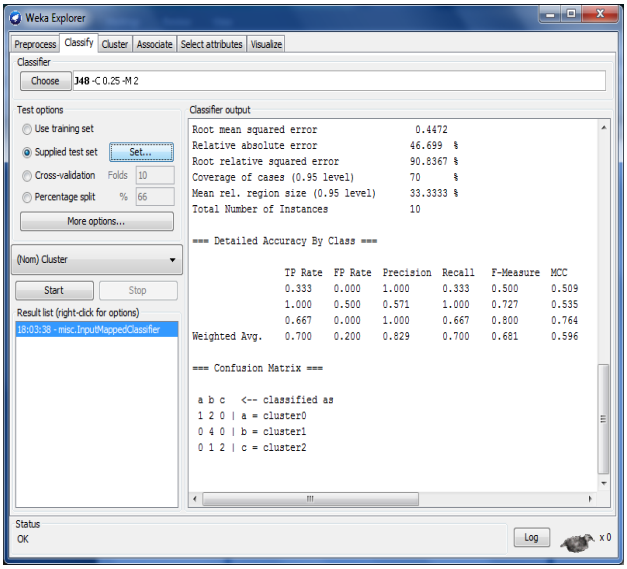


Figure 6: Result after applying J48 algorithm

6. Conclusion

This work gives the result after analysis to distribute the risk in different passport offices used in this work where the sensitive data originates. Among the five offices that are considered in this work select the one which is dominant in the high risk and in low risk, this is done using clustering approach. This technique gives the result that office C has high risk and office D has low risk. Then these clusters are further used for the validation, which validate the data with the help of classifying algorithm. This gives correct and incorrect instances as a result. So, this work essentially helps in identifying the fraud entries in the database. This research work also provide the information about the passport issuing offices where the sensitive data originates and validation part provide information about the incorrectly placed instances and correctly placed instances.

7. Future Work

There is always a scope of improvement in any research and so is with this work also. This work used primary database, it is a kind of prototype system using this real world problems were analyzed in future. We can also profile Risk categories for each and every passport issuing centre. Hybrid techniques can also be used in future work to improve the accuracy of the prediction of the Risk Score and demographic profiling for passport issue from each centre.

8. Acknowledgement

Authors would like to thanks to their head Dr. Rajan Vohra, Head of Department of CSE & I.T department, PDMCE, Bahadurgarh, India for his valuable support and help.

References

[1] SuchetaGulia, Dr. Rajan Vohra, Minakshi, Gimpy, "Risk Computation and Identification in Passport Data Analysis" in communication.
[2] Dr. SankarRajagopal, "Customer Data Clustering Using Data Mining Technique" International Journal of Database Management Systems (IJDMS) Vol.3, No.4, November 2011.
[3] Osmar R. Zaïane,"Principles of Knowledge Discovery in Databases - Introduction to Data Mining", CMPUT690, 1999.
[4] Chakrabarti, S., 2002: Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kauffman.
[5] Warrender, C., S. Forrest, and B. Pearlmutter, 1999: Detecting intrusions using system calls: Alternative data models. IEEE Symposium on Security and Privacy.
[6] Rabiner, L. and B.-H. Juang, 1993: Fundamentals of Speech Recognition, Prentice-Hall, chapter 6.
[7] Durbin, R., S. Eddy, A. Krogh, and G. Mitchison, 1998: Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press
[8] Han, J. and M. Kamber, 2000: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.
[9] Terry Harris, (2008)"Optimization creates lean green supply chains", Data Mining Book
[10] Matt Hartely (2005)"Using Data Mining to predict inventory levels" Data Mining Book
[11] Hu, Tung-Lai, &Sheub, Jiuh-Biing (2003). "A fuzzy-based customer classification method for demand-responsive logistical distribution operations"