Complex & Intelligent Systems (2023) 9:3681–3694
https://doi.org/10.1007/s40747-021-00548-x

ORIGINAL ARTICLE

Integration and classification approach based on probabilistic semantic association for big data

Vishnu VandanaKolisetty[1] · Dharmendra Singh Rajput[2]

Received: 30 March 2021 / Accepted: 17 September 2021 / Published online: 11 October 2021

Abstract

The process of integration through classification provides a unified representation of diverse data sources in Big data. The main challenges of big data analysis are due to the various granularities, irreconcilable data models, and multipart interdependencies between data content. Previously designed models were facing problems in integrating and analyzing big data due to highly complex and dynamic multi-source and heterogeneous information variation and also in processing and classifying the association among the attributes in a schema. In this paper, we propose an integration and classification approach through designing a Probabilistic Semantic Association (PSA) method to generate the feature pattern for the sources of big data. The PSA approach is trained to understand the data association and dependency pattern between the data class and incoming data to map the data objects accurately. It initially builds a data integration mechanism by transforming data into structured and learn to utilize the trained knowledge to classify the probabilistic association among the data and knowledge patterns. Later it builds a data analysis mechanism to analyze the mapped data through PSA to evaluate the integration efficiency. An experimental evaluation is performed over a real-time crime dataset generated from multiple locations having various events classes. The analysis of results confined that the utilization of knowledge patterns of accurate classification to enhance the integration of multiple source data is appropriate. The measure of precision, recall, fall-out rate, and F-measure approve the efficiency of the proposed PSA method. Even in comparison with the state-of-art classification method and with SC-LDA algorithm shows an improvisation in the prediction accuracy and enhance the data integration.

Keywords Integration · Probabilistic association · Semantic classification · Big data

Introduction

Big data have started to extend everywhere and has been utilized in many fields such as computer vision, machine learning, financial, and social analytics. Traditional information integration systems are systems based on a complex design that connects limited sources, is relatively durable, and is usually time-consuming. On the other hand, data applications are becoming more and more widespread and require flexibility and uncertainty. Applications that involve multiple sources, such as Google Base or in-depth web or tools require the automatic deletion of the semantic match between the intermediation schema and the data sources, which can be approximated. Such as, in [1] a framework for learning the records of pharmacy automatically and in [2] a mechanism to simplify the complex traffic information through autonomous coordinated control in big data.

In the vast amount of information available in big data, we are often concerned about the vast and complex databases like Facebook, Twitter, and LinkedIn, which nearly one million active users store, share, and connect around the world every month [3]. It is necessary to capture, store, transfer, search, share, analyze and efficiently process the data of this very large database [4]. It is not possible to collect and combine such large amounts of information and knowledge using traditional methods and tools. In [5] the analytical study of the multidimensional structured and unstructured big data, and in [6] the mechanisms of integration of big data are discussed. Thus, an automated system is required to analyze this text to understand and analyze the use of different applications.

✉ Dharmendra Singh Rajput
dharmendrasingh@vit.ac.in

Vishnu VandanaKolisetty
kvishnu.vandana2016@vitstudent.ac.in

[1] SCOPE, Vellore Institute of Technology, Vellore 632014, India

[2] SITE, Vellore Institute of Technology, Vellore 632014, India

The mechanism of data integration and classification will provide a unified platform for collective associating multiple sources of data and also enhance the capability to exchange information between various information-sharing systems as in [7] the data integration with the interaction with schema and data records relation, and in [8] the perspective of data association for integration with machine learning studies are presented. Existing data integration systems are mainly extensions of traditional databases and are structured and data can be designed utilizing a few of the conventional data models [9]. In addition, the systems of data integration are well aware of the procedure for the data sources are to be adapted for the association scheme using an appropriate data integration system [10].

Often the collected data will not be in one format or ready for analysis. Every time events happen, it is necessary to follow the development of events. In general, the need to follow events has a strict time limit; in particular, the event is a disaster or an emergency. Therefore, the identification of events [11] and temporary data [12] is important to control the development of events, including the prevention of disasters or the further improvement of outcomes. After some emergencies, it is important to gather enough information about emergencies as soon as possible and to conduct data analysts to support subsequent decision-making [13]. Historically related events can be utilized to predict and provoke the growth of existing incidents. It cannot leave the data in this form and still analyze it effectively. On the contrary, it requires a data extraction process that extracts the necessary information from the main sources and expresses it in the form of a structure suitable for analysis [14].

Problem

The most available data in the big data distribution are either in the form of structured or unstructured. It is essential to process and simplify them to extract valuable information for analysis. The structured data can be readily utilized for the information analysis [15], as it provides the basis of the data required for the quantification and analysis utilizing the structure properties and its values. But, in the case of unstructured data, it is essential to convert it into structure form through utilizing the Natural language processing (NLP) process primarily. To extract the information required, the basic process of cleansing and data transformation is needed to be performed [16].

Unfortunately, it can be very difficult to determine the accuracy of semantic maps. Manual schema compatibility is generally time-consuming and cannot be implemented, especially with large databases [17]. Since the semantics of schema objects cannot be fully derived from data and metadata, automatic schema matching is essentially uncertain [18]. In our new approach, the uncertainty in the defined maps is represented during the scheme adaptation process; the uncertainty schema is spread over the merging process and is described in the resulting integration schema.

The main issue with integration and analyzing big data is the complex interplay between the granulation of variable data, inconsistent data models, and data content. It was therefore not easy to find an effective knowledge model and it was very difficult to manage and group the relationships between the data attributes. Existing integration and analytical frameworks like pharmacy data record matching framework [1], biodiversity data retrieval framework [4] and, data records with schema matching framework [7] are facing many challenges in analyzing this large-scale data due to the complexity of the distribution of data and the proliferation of multi-source data. To solve this problem, and integration and classification model based on the Probability Semantic Association (PSA) of the attribute and generation of knowledge pattern is proposed for a large data source.

Contribution

The objective of this paper is to propose a new approach for data integration and classification for uncertain and unstructured data through a Probabilistic Semantic Association (PSA). It aims to simplify the core problem of data integration by identifying the corresponding data objects to their matching class schema. Information about semantic maps is important to create an integrated scheme [19]. Recent studies [20] have shown that if all semantic adaptations are known, the schema combination can be performed semi-automatically. The PSA approach utilizes the data schema of the Hadoop Framework to build the integration and analysis mechanism. It modifies the most common probabilistic algorithm Naive Bayes (NB) to constructs the required knowledge patterns, which is utilized for classifying the anonymous data automatically to map to their matching schema.

The illustration model of the proposed PSA is shown in Fig. 1. It states that the crime data occurring at different locations need to be mapped to their respective classes through learning the crime post from location L_1 to L_d.

Most of the available techniques such as records to schema matching [17] and schema integration through semantic mapping [19] try to define a single mapping for each pair of objects, which of course can be wrong. However, in general, it is not possible to fully automatically identify correct semantic maps where the schemes are almost identical, and there are subtle differences in the semantics of schema objects that make it very difficult to discover true semantic maps. Thus, an automated data matching technique can create an equivalence map between two criminal entities in C_1 and C_2, based on semantic comparison to minimize the process and time load. Using a possible association, our approach allows us to include potential matches and assign probabilistic values to

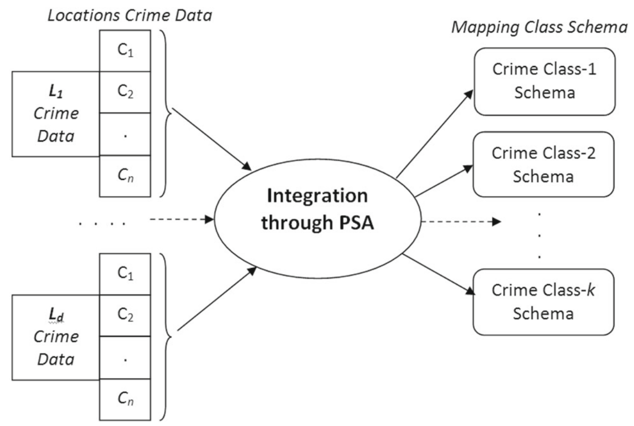


Fig. 1 Integration model using PSA

them. This significantly reduces the cost of circuit integration, allowing it to be fully automated and thus expanded to a large number of data sources.

The contributions of this paper are as follows:

- Presents an integration and classification model for associating big data through probabilistic semantic association. The proposed model includes several steps of the mechanism
- The first step, data integration which allows data preprocessing of the data and performs data extraction and transformations from different sources to build an event class knowledge pattern with learning the partial integration of the schemas entity relationship.
- The second step explores data integration analysis using the PSA approach. Later, we analyze the outputs with NB probabilistic prediction, and also with few Ensemble classifiers to conclude the enhancement of the proposal.

In section-II, we discuss the background study related to big data integration and semantic association, section-III, we explain the proposed integration and classification model, section-IV explains experiment evaluation and section-V presents the conclusion of the paper.

Background study

In terms of big data management and analysis, traditional databases face many challenges when managing multiple sources with similar or different data, because the big data associated with patterns is difficult and dynamic. The types and relationships of resources can grow over time and it is not possible to change the location of the database frequently due to the strong connection between the database and the programs [21].

According to Sun et al. [15], introduce a semantic-based structural similarity for the first time and then propose an approach to measure the semantic-based structural similarity between networks with the computing theory for semantic relations as the foundation. They also aim at exploring the algebraic computing approach of semantic relations [22]. Even Sun et al. [23] exploring an extensible and active semantic information organization model for IoT to meet the requirements, and the primary idea is "Object-cored organizing data, event-based explaining data, and knowledge-based using data". Lee et al. [24] conducted an experiment that collected repeated similarity measures for each pair of documents in a small corpus of short news documents using word-based, n-gram, and Latent Semantic Analysis (LSA) approaches. Zhang et al. [25] work on text-based event temporal resolution and reasoning including identification of events, temporal information resolutions of events using the rule-based temporal relation reasoning between events and relevant temporal representations. Agichtein et al. [27] mine tables present in data warehouses and relational databases to develop an automatic segmentation system to overcome limitations of existing supervised text segmentation approaches, which require comprehensive manually labeled training data for classification. Kumaran et al. [28] show how performance on New Event Detection (NED) can be improved by the use of text classification techniques as well as by using named entities in a new way. Dalvi et al. [29] describe the foundations of managing data where the uncertainties are quantified as probabilities. Bovenzi et al. [34] present a novel hierarchical framework for traffic classification of anonymity tools enabled by big data-paradigm and capitalize the machine learning. Liu et al. [36] propose BRBCS-PAES: a selective ensemble learning approach for BRB Classification Systems (BRBCS) based on Pareto-Archived Evolutionary Strategy (PAES) multi-objective optimization. Zhang et al. [37] a novel multi-gram convolution neural network-based self-attention model with a recurrent neural network framework. Reddy et al. [38] present an analysis of ML algorithms-based classifier's performance on Big data utilizing dimensionality reduction techniques. It suggests a great extent of improvisation in classification using Principal Component Analysis (PCA) and Linear Discriminate Analysis (LDA) techniques.

In essence, traditional databases are supposed to remain stable and reliable for long-term change, but not for data analysis that focuses more on interpersonal relationships across different sectors of the data storage [9]. It can be used as hierarchical models (such as tree models) to manage data for data analysis, and network models to organize multiple sources and identical data. Hierarchical and network models can reflect complex relationships between individual data objects [15]. However, domain-specific applications are required for the semantics of fragments of individuals and relationships. Therefore, a new data model or data orga-

Table 1 Related works for the solution of big data integration and classification

R. no	Work	Objective	Method
[1]	Design framework to deploy well-known rules to match records for the integration from different databases	Data integration	Machine learning and pattern detection
[2]	An autonomous traffic management mechanism by integrating the information	Data integration and control	Multi-agent CPS system
[4]	Design a computational framework to integrate and manage biodiversity data	Data Integration, Information Retrieval	Biodiversity data management
[5]	Present analytical study on unstructured and multidimensional big data for event information extraction	Learning Big data information extraction techniques	Learning-based techniques
[7]	Design a Schema Matching and Record Matching rules for improving data integration	Data Integration	Schema and record matching
[9]	A deep learning through semantic analysis of texts entity for recognition in Health documents	Relation extraction	Deep learning
[10]	Design a data integration system to incorporate uncertain measures in a probabilistic framework	Data Integration	Probabilistic features relationship
[11]	An online content processing and classification for automatically detecting suicide events	Text classification	Machine learning, Feature selection
[12]	Design a multi-information location data fusion system for information sharing	Heterogeneous data integration	Information physical fusion system (CPS)

nization model for managing and analyzing big data needs to be explored [22]. Some semantic data models have been suggested to provide multiple sources and identical data features over big data. Information based on semantically organized data or examples can help analysts find the mind to make decisions [23]. It is important to store a large amount of data for use and solve problems caused by computer constraints. On the one hand, big data sizes are big and data analytics need to respond quickly. Fortunately, cloud computing technologies have allowed us to process and analyze mass data to gain knowledge and insight [24]. A few of the related works which provide solutions for the data integration and classification are summarized given in Table 1.

To utilize big data, we are interested in events or knowledge of big data, not the details of large data. This event is a collection of definitions that include phenomenon-forming elements such as information hidden in large data, when, where, who, why, how, and information hidden in large data [25]. Events are happening every day and every time in the world. So, different events shape a person's history and memory. Many of these events are intimately related to our daily lives and work and several significant events will have a negative impact on us in the future. Even the security and privacy aspects of big data management are a major challenge. Deepa et al. [39] present the different data security challenges and complexity in big data analysis and management. Tang et al. [40] discuss the threats and vulnerability in big data through unveiling the multivariate dependencies arrangement amongst various vulnerability risks. Khan et al. [41] suggest a decision tree algorithm to predict the most relevant and irrelevant features for effectively detects security threats using machine learning classifiers.

This event is closely related to semantics, and languages are one of the most important means of demonstrating semantics. Thus the phenomenon of logic has become part of the theory of semantics. The semantics of incidents have their roots in Davidson's suggestion that verbs have an additional 'hidden' place for arguments [26]. This idea has proved very fruitful in the semantics of natural language with a wide range of consequences. Moreover, the texts are expressed in different languages, so the automatic extraction of events and the expression of the semantics of accidents are close to the development and understanding of natural language [27]. The role of events in information analysis has a significant role [28]. To understand the situation and make the right decision, it is necessary to clarify the sequence of events, and important sources of information are organized according to the time sequence of events. During the development of the event, the people associated with it will do something elsewhere. Many events combined have formed parts of history. So analysts urgently need sources of information organized into various events.

In the case of an incident-based information organization, time is a natural guide to planning information in the event of an incident. However, the incident does not have a coherent explanation. For example, in language research, an event could be a verb or a noun. In industrial control, the event can be a state change. Like, the events of activities being occurs in a crops plants which can be cause for diseases is will be significant to classifying towards improvisation crop [42]. The definition of events is different, but each event is

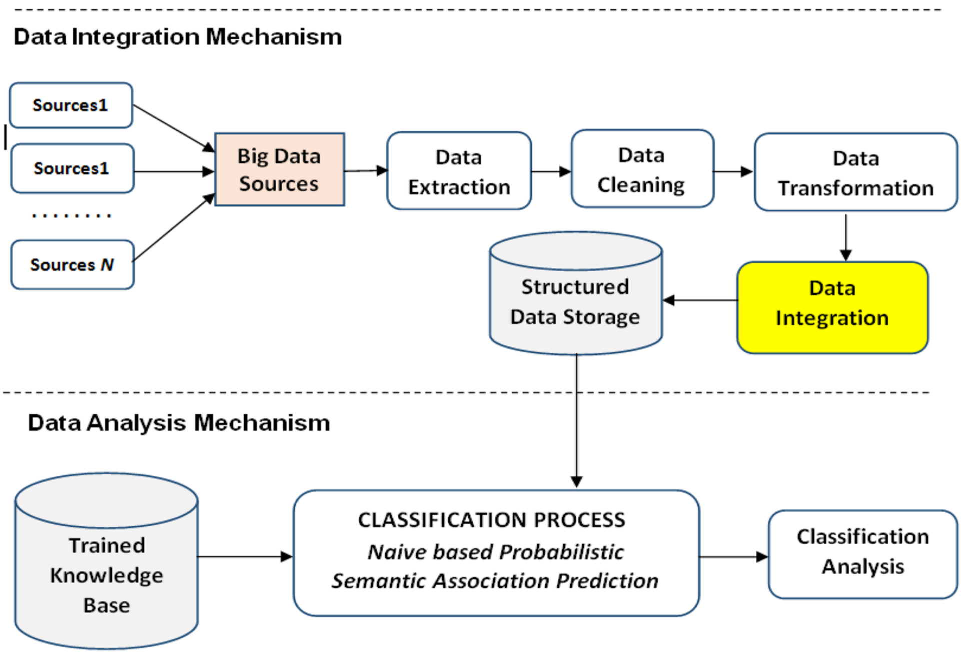


Fig. 2 Integration and Classification Model

essential for two key elements or attributes: time and location. The time and geographic information of citing and analyzing events are everywhere. Likewise, time is an essential element of any information space. Activities are not only close to the semantic organization of information resources, but also have different types in different intellectual applications.

Event and time information is usually used to search for information and answer questions. The large-scale interactions between people on the Internet make it difficult to study interdisciplinary theories describing collaboration [29], which are interconnected to some extent due to their nature.

Our proposed approach is also based on an incident-based dataset analysis which is discussed in detail in the next section. We use existing machine learning and data extraction techniques to facilitate the rich analysis of integrated data tables [30].

Proposed integrated and classification model

The proposed approach consists of two-phase functionality, Integration of large data sources into structured data, and data classification for analysis as shown in Fig. 2. In the first stage, large data sources are integrated with structured data by executing a mechanism for pre-processing for diverse data sources to be utilized for learning and analysis, and in the second stage, the classification and forecasting of structured data are carried out for investigation.

The approach is evaluated over a big dataset published by SFPD [31]. The database is located in a separate location through the Criminal Reporting System. Predicts events based on activist data, which will be important for case monitoring and analysis. To facilitate this forecast analysis on datasets; we perform a PSA interpretation between the event descriptions taken to predict the event classes. In the following section, we discuss the methodology of the Data integration mechanism and Data Analysis Mechanism.

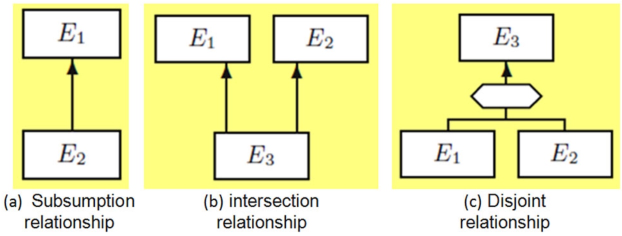
Data integration mechanism

Big data are an uncertain source of information recorded from different streams and environments. There may be some very uninteresting information that needs to be filtered and cleaned before the recorded data can be analyzed. The data integration mechanism implements "data extraction" and "data cleaning" methods for constructing structured data. The collected data are not ready for analysis. It tends to be incomplete, noisy, and inappropriate. Data cleaning (or data cleaning) routines try to fill in missing values, correct noise when identifying outsiders, and correct inconsistencies in information.

Data extraction and transformation

This method implements data extraction through reading different data sources, which are typically integrated from a different source for creating a unified structure. It also updates or uploads standard information from sources. It performs data extraction generally using the app interface. The approach utilizes a Query Editor of Hadoop interface to mine and install a database schema.

The process of data cleaning and transformation fills the missing values for the attributes and removes the irrelevant



(a) Subsumption relationship (b) intersection relationship (c) Disjoint relationship

Fig. 3 Partial Integrated Schemas of ER Entity Relationship

tuples employing an attribute construction method. It detects errors or duplicate tuples in the data and rectifies them wherever is possible. The resulted from cleaned data will be effective for integration.

Process of data integration

In data integration, we merge the cleaned data from multiple sources into a coherent data store. The sources of information included in this approach are numerous CSV documents. The data is stored in a form designed for analysis. The process of integration implements data mapping to their respective schema class. A set of classes needed for the mapping is identified before correlate the association process for mapping. Each class is described by a set of attributes values which can map to relate each class is constructed through a training process to have a pattern knowledge base.

The construction sample is carried out on the storage of knowledge base transformation data [32]. We used 40% of the historical data to create a sample database for each event class category, and the remaining and final data of events can be classified to forecast using a sample database that forms the basis of data analysis and decision making. The Training Construction Module conducts individualized information training to build a design that can be used for real-time streaming of information and is used to assess the appeal of the patterns that occur.

In [32] it is defined as the compilation of plans based on the mapping definitions between the items. Standard rules for generating two-dimensional project changes [9] and merging two schemas have been translated. Implementing three such rules on E_1 and E_2 components is shown in Fig. 3.

Figure 3a shows a partially coordinated plan that is created when the subsumption relationship between the two “ER” topics is described. Figure 3b shows a partially integrated schema created by the intersection between two entities is identified, and Fig. 3c shows a partially integrated schema created by the interaction between two entities based disjoint relationship.

This process of knowledge pattern generation is a supervised learning process through understanding the relationship [9] and learning the pattern matching rules for information extraction [14] utilized as the class label of each training tuples is offered. Each class reveals interesting information patterns hidden in large training data sets that are represented as useful information for class label patterns. These knowledge patterns act as relation attributes for the data entities to perform the integration process as per the above illustration in Fig. 3.

As described in Algorithm-1, we often collect items to construct related values in trained data sets. The construction of a knowledge sample is often accomplished by assembling the product by studying the related values in the taught data sets, as shown in Algorithm 1.

Algorithm-1: Building of event class knowledge pattern

```
Input: $E_{Class}$ [ ] → Set of event class.
       $T_{Data}$ [ ] → Dataset for training.

Output: $CK_p$ → Each class knowledge pattern.

Init $x=0, y=0, z=0;$
$s = sizeOf(E_{Class});$

Loop1: while $(x < s)$
    //-- Get event class from event set
    $E = E_{Class}[x];$

    //-- Get all data records having class E.
    $R_{Data}[\,] = T_{Data}[E];$
    $d = sizeOf(R_{Data}[\,]);$
    $i = 0;$
    Loop2: while $(y < d)$
        //-- Get data records at t index from $R_{Data}$.
        $V = R_{Data}[y];$

        //-- Build data record terms set.
        $W[\,] = Split\text{-}Record(V, \text{"space-delim"});$
        $t = sizeOf(W[\,]);$

        Loop3: while $(z < t)$
            $term\text{-}count = 0;$
            $q = W[z];$
            //-- Get the term frequency of the term q in W [ ].
            $term\text{-}count = \text{getTermCount}(q, W[\,]);$
            if $(term\text{-}count > 1)$ {
                $K_p[i] = q;$
                $i{+}{+};$
            }
            $z{+}{+};$
        Loop3: End while
        $y{+}{+};$
    Loop2: End while
    $CK_p[x] = K_p;$
    $x{+}{+};$
Loop1: End while
    return $CK_p;$
```

The established knowledge pattern, CK_p will be the basis for the prediction of the present event, and also the incoming new data will be the references of the analysis to the current event. We modified an NB algorithm by integrating the probability of a semantic combination between knowledge patterns and test data for classification and prediction [29].

The identification of the above semantic relationships during schema matching is done by two-way comparison. Based on the design architecture, which is consisting of knowledge patterns for data objects feature matching will be utilized with the different types of data to compare statistical data of the schema objects. Knowledge patterns generate degrees of similarity, are then generalized, and the semantic relationships between schema objects and data records are shown using user-defined thresholds.

Data analysis using the PSA approach

The classification process as a whole created a great opportunity for the class to predict using the probability of uncertainty. This approach proposed a PSA approach for automates the incident prediction. PSA creates semantic maps based on the choice of a probabilistic association that determines the relationship between the established knowledge base and current data.

To illustrate the mechanism of PSA, let's assume, Z_{Data} is a set of test event data having n data records, CK_p is the set of knowledge patterns generated having k number of patterns and E_{Class} is the collection of m events classes. So, to map the n data records of Z_{Data} with the E_{Class} class events, we predict the association probability using the NB method for each data record of Z_{Data} with each k pattern of CK_p as present in the PSA approach in Algorithm 2.

Algorithm-2: PSA Approach

```
Input: $E_{Class}$ [ ] → Set of event class.
       $Z_{Data}$ [ ] → Dataset for testing.
       $CK_p$ [ ] → Event Class knowledge patterns.

Output: Mapped event class C of each data in $Z_{Data}$

Init x=0, y=0, z=0;
t = sizeOf ($Z_{Data}$);

Loop1: while (x < t)
    //-- Get data records at t index from $Z_{Data}$.
    V = $Z_{Data}$ [ t ];
    s = sizeOf ($E_{Class}$);

    //--Comparing with event classes
    Loop2: while (y < s)
        freq_asso_cnt=0;
        //-- Read event class knowledge patterns
        E [ ] = $CK_p$ [ y ];

        //-- Build data record terms set.
        W [ ] = Split-Record (V, "space-delim");
        m= sizeOf (W[ ] );

        Loop3: while (z < m)
            q = W [ z ];
            //-- Get the frequency of the term association q in E [ ].
            fcnt =getTermCount (q, E [ ] );
            if (fcnt > 0) {
                freq_asso_cnt++;
            }
        Loop3: End while

        Event_class_asso[ y ] = freq_asso_cnt;
      y++;
    Loop2: End while

    //-- Get top 3 highest frequency association count class from Event_class_asso[ ].
    H [ ] = getHighestClass_ Asso( 3, Event_class_asso[ ], $E_{Class}$ [ ] );

    //-- Probabilistic Semantic Association
    n = 0; z= 0;
    Loop4: while (n < sizeOf (H [ ]) )
        //-- Build data record terms set.
        W [ ] = Split-Record (V, "space-delim");
        m= sizeOf (W[ ] );

        psa_val= 0;
        Loop5: while (z < m)
            q = W [ z ];
            psa_val = $\Sigma$ q $\epsilon$ E [ ];
             psa_tot=psa_tot+psa_val;
            z++;
        Loop5: End while
        psa_tot_assoc [n] = psa _tot;
        n++;
    Loop4: End while
    //-- Get the class having highest PSA association value for the data record V.
    C = getHighestPSA_Class ( psa_tot_assoc [ ], H [ ] );
    //-- Data record V mapped to event class C.
    V → C;
    t ++;
Loop1: End while
```

Table 2 Event class knowledge patterns

P.Id	Event class	Generted knowledge patterns
1	Assault	Aggravated, Assault, Police, Officer, Sniping, Bodily, Force, Gun, Battery, Former, Spouse, Dating, Relationship, Child, Inflicting, Injury, Traumatic, Condition, Firearm, Discharging, Occupied, Bldg, Vehicle, Aircraft, Grossly, Negligent, Manner, Resisting, Peace, Causing, Serious, Death, Knife, Stalking, Threat, Resist, Executive, Threatening, Threats, Against, Life, School, Teachers, Unlawful, Dissuading, Threatening, Witness, Willful, Cruelty
2	Bribery	Dissuading, Witness, Victim, Bribery, Executive, Officer, Witnesses
3	Drug, Narcotic	Hypodermic, Needle, Syringe, Possession, Prescription, Forge, Alter, Encourage, Minor, Amphetamine, Barbituates, Controlled, Substance, Hallucinogenic, Heroin, Meth, Opiates, Encouraging, Cocaine, Marijuana, Failure, Register, Narcotics, Addict, Furnishing, Loitering, Methadone, Paraphernalia, Opium, Derivative, Base, Schoolyard, Trafficking, Influence, Drugs
3	Disorderly conduct	Disturbing, Peace, Commotion, Fighting, Swearing, Committing, Public, Nuisance, Maintaining, After, Notification
5	Embezzlement	Embezzlement, Grand, Theft, Brookers, Agents, Collector, Employee, Property, Carrier, Leased, Private, Public, Private, Official, Petty, Embezzled, Vehicle
6	Family offenses	Children, Abandonment, Neglect, Child, Concealment, Removal, Without, Consent, Desertion, Failure, Provide, Parents, Immoral, Acts, Drunk, Presence, Minor, Proper, Parental, Care
7	Forgery, Counterfeiting	Checks, Legal, Instruments, Uttering, Forged, Forgery, (Felony), (Misdemeanor), Make, Pass, Fictitious, Possession, Intent, Counterfeiting, Coins, Notes, Plates, Dies, Drivers, License, Id, Card, False, Entries, Records, Returns, Government, Corporate, Seals, Identification, Possess, Sell, Money
8	Fraud	Access, Card, Information, Theft, Credit, Incomplete, Counterfeit, Use, False, Claims, Presenting, Government, Pretenses, Grand, Petty, Forgery, Privacy, Invasion, Short, Change, Trick, Device, Attempted, Defrauding, Innkeeper, Vehicle, Repairman, Id, Fraudulent, Auction, Application, Automated, Teller, Impersonating, Police, Putting, Slugs, Coin, Operated, Machines, Parking, Meters
9	Kidnapping	Attempted, Kidnapping, Adult, Victim, Juvenile, Child, Stealing, False, Imprisonment, During, Robbery
10	Vehicle theft	Vehicle, Recovered, Auto, Bus, Camper, House, Car, Motor, Home, Mobile, Trailer, Motorcycle, Other, Attempted, Stolen, Outside, Automobile, Miscellaneous, Truck, Tampering, With

Table 3 Accuracy of mapping of PSA

No. of data records	# of correctly mapped	# of incorrectly mapped	# of not mapped	Mapping (%)	Accuracy (%)	Error (%)
1000	990	5	5	99.50	99.00	1.00
2000	1906	82	12	99.40	95.30	4.70
3000	2841	124	35	98.83	94.70	5.30
4000	3819	181	0	100.00	95.48	4.53
5000	4682	209	109	97.82	93.64	6.36
6000	5599	363	38	99.37	93.32	6.68
7000	6584	228	188	97.31	94.06	5.94
8000	7483	480	37	99.54	93.54	6.46
9000	8541	315	144	98.40	94.90	5.10
10,000	9289	603	108	98.92	92.89	7.11

The mechanism of PSA estimates the relation with the event classes C, for each data record of Z. It will simplify the classification of the data which is uncertain and complex in matching to the event class through the PSA algorithm.

As per the above discussions, the proposed PSA is sensitive to two parameters as the frequency of term association and event class knowledge patterns (CK_p []). The closeness of association of an event to a class is depending on the highest PSA association value. As the precise computation of PSA association value is highly depends on the frequency of term selection, so the sensitivity of terms selection is very significant to build the CK_p. The creation of the knowledge pattern of CK_p of an event class using Algorithm 2 is vary based on the frequency of term count which can be tuned to have a variety of patterns of events. The preciseness in the creation of CK_p results in the enhancement of the classification precision while mapping.

Let's consider a dataset having 10 crime events classes. The events that have more than one attribute of frequency are categorized to form a pattern. We consider the 10 event class to build the event class knowledge patterns CK_p for each data record in the datasets as given in Table 2.

Now utilizing these generated patterns we classify the records of the dataset to their respective mapping class through predicting their mapping event class is analyzed in Table 3 as given below. It shows that an average of 98.91% of mapping can perform using the knowledge pattern, where it achieves an average of 94.68% accuracy with an average of 5.32% of error in mapping.

The analysis of the PSA was performed over distributed crime event dataset SFPD is briefly discussed in section "Experimental evaluation".

Experimental evaluation

To evaluate the proposed PSA approach we collect the dataset of SFPD open data released [31] for a period of 12 years from 2003 to 2015 of San Francisco districts. It presents 9 data fields and more than 1.5 million data records for each incident as discussed below.

- Dates: date and time when the crime incident happened
- Category: Category of the crime incident having 39 different categories of events. This is the objective variable that will be predicted by PSA.
- Description: Detailed description of the crime incident.
- DayOfWeek: the day of the week
- PdDistrict: name of the Police Department District
- Resolution: how the crime incident was resolved
- Address: the approximate street address of the crime incident
- X: Longitude
- Y: Latitude

Table 4 Event class and description data

Event class	Event description
Assault	Threats against life
Assault	Aggravated assault with a deadly weapon
Drunkenness	Under influence of alcohol in a public place
Larceny/Theft	Grand theft from unlocked auto
Other offenses	Traffic violation arrest
Secondary codes	Domestic violence
Suspicious occ	Investigative detention
Vehicle theft	Stolen automobile
Warrants	Enroute to outside jurisdiction
Warrants	Warrant arrest

Table 5 Prediction parameters

TP (True positive)	No. of event data correctly predicted
TP (True negative)	No. of event data incorrectly predicted
FN (False negative)	No. of event data not predicted

We utilize the Category information data field name as "Event Class" and Description as "Event description". The event class provides the kinds of crime incidents identified and the description defines the events. It has collectively 39 different crime events, where few of them are described in Table 4.

Implementation and evaluation measures

We manually collect and extract data sets from [31] for application and evaluation, and it is eliminating data noisy and errors through performing data cleaning. The process data is loaded into a standard schema of the Hadoop Framework. An "*sfpd_db*" database schema is designed to transfer data from a standard schema to an "*sfpd_db*" database utilizing a SQL script.

To construct the knowledge base pattern we considered 40% of the data records. A designed program developed using java is developed to process the data and creates the pattern of knowledge for each class. To measure the effectiveness of the proposal for testing, we are working on different data records from the traditional NB and the proposed PSA classification method.

It measures the Precision, Recall, Fall-out rate, and F-measure metrics to evaluate the outcome results of the datasets. These measures are calculated based on the values of the confusion matrix determined during the evaluation process considering the observation as given in Table 5.

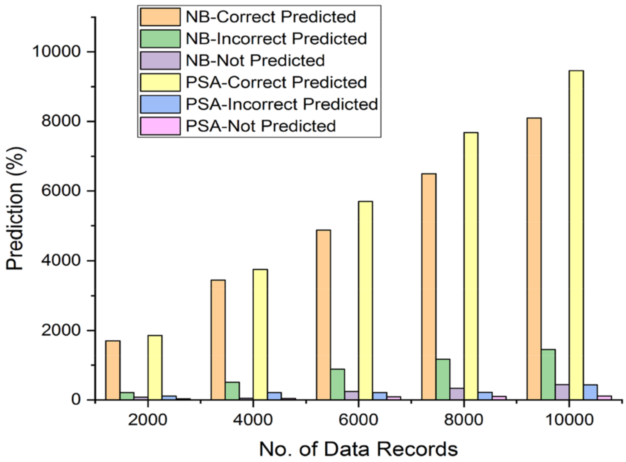


Fig. 4 Prediction Comparison between Naïve Bayes and PSA

Based on the observation we compute the following equations,

$$\text{Pr}\,ecision\ (P)\ =\ \frac{TP}{TP+TN} \tag{1}$$

$$\text{Recall}\ (R)\ =\ \frac{TP}{TP+TN+FN} \tag{2}$$

$$Fall\ Out\_Rate\ =\ \frac{TN}{TP+TN} \tag{3}$$

$$F-Measure\ =2\times\frac{(P\ \times R)}{(P+R)} \tag{4}$$

$$Accuracy\ (ACC)\ =\ \frac{TP+TN}{TP+TN+FN+FP} \tag{5}$$

The evaluated results are explained in the next section against a different number of records classified.

Results

We perform the classification and prediction of datasets, using traditional Naïve Bayes and PSA to measure the improvisation in comparison Naïve Bayes classification is more commonly used in text classification and prediction tasks.

Comparison analysis of PSA with Naïve Bayes prediction

Predictive comparisons of true, false, and unpredictable outcomes are given in Fig. 4. The results of PSA indicate that the Naïve Bayes is improvising on incorrect prediction.

To measure the positivity and sensitivity prediction we measure precision and recall rate. Figures 5 and 6 presents the precision and recall rate comparison. In comparison to Naïve Bayes, the proposed PSA shows an improvisation of an average of 10% in the accuracy rate and recall rate. Improvisation is associated with enhanced computation via PSA.

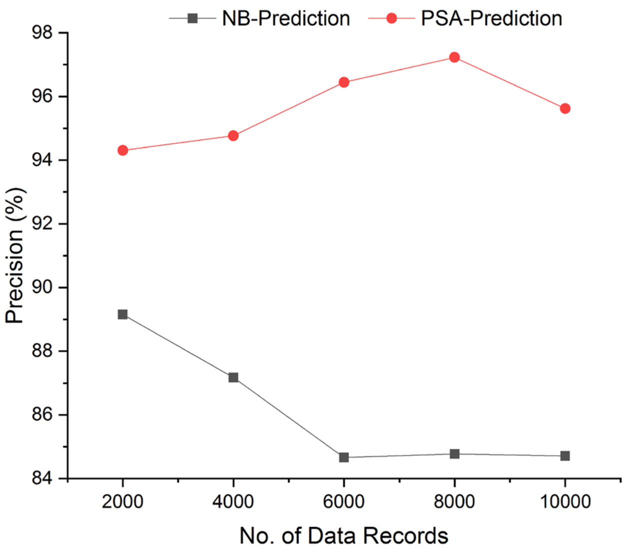


Fig. 5 Precision Rate comparison between Naïve Bayes and PSA

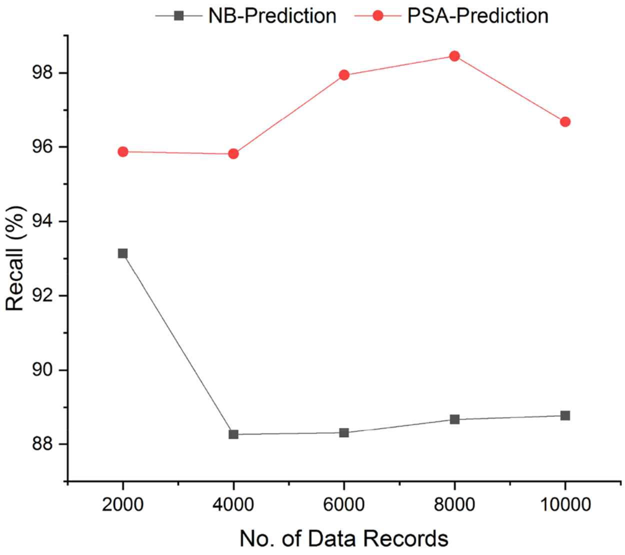


Fig. 6 Recall Rate comparison between Naïve Bayes and PSA

Figures 7 and 8 present the fall-out and F-measure rate. The fall-out rate measures the rate of incorrect prediction against a predicted total data entry. PSA shows an average drop rate of 9% compared to Naïve Bayes. The F-measure that combines precision and recall to measure harmonic means appears to be enhanced in comparison to Naïve Bayes classification. According to the observation results of all the above measures, the effectiveness and applicability of the proposal in feature analysis and decision-making tasks are proved.

The effectiveness of the above result measures is due to the accuracy of mapping the required class of events for the incoming events information at runtime. An attribute corre-

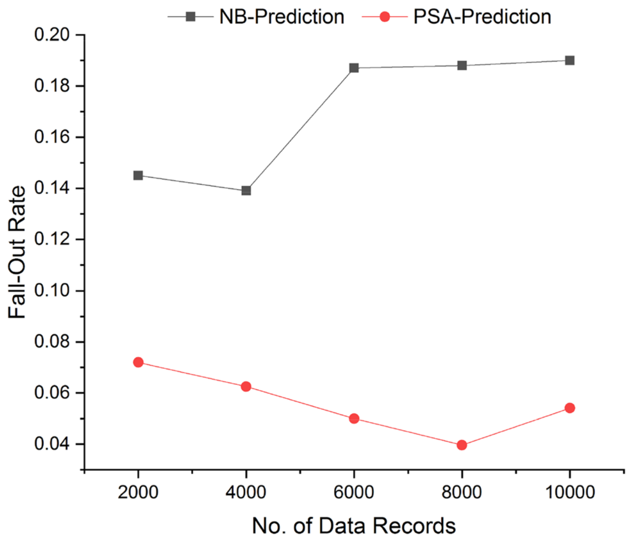


Fig. 7 Fall-out Rate comparison between Naïve Bayes and PSA

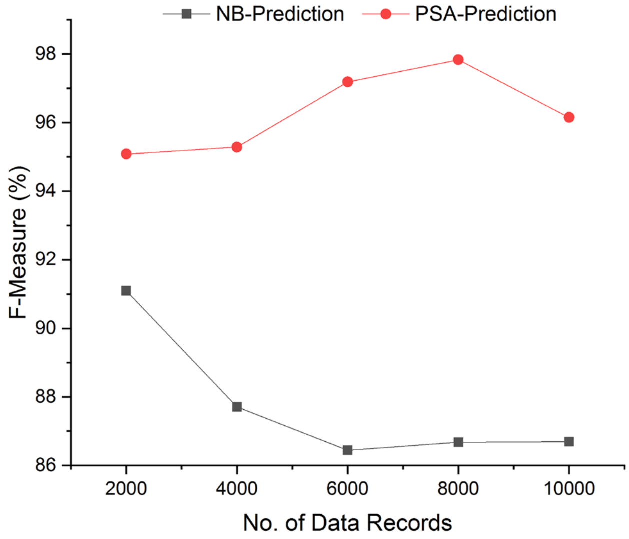


Fig. 8 F-Measure comparison between Naïve Bayes and PSA

spondence is of the event form as $w_{a,b} = (s_a, t_b)$, where s_a is a source attribute in the schema S and t_b is a target attribute in the schema T. The entire event schema refers to the first interpretation over the knowledge patterns and the second one as the event class of probabilistic mappings. In varying the number of data records from 2000 to 10,000 the probabilistic mapping of the events concerning events class tuples defines a relationship between instances of S and instances of T that are consistent with the mapping.

The algorithm of PSA with a set of event class, event class knowledge patterns and a set of testing data behaves differently with varying data records 2000 to 10,000. In case of the lowest 2000 input the algorithm mapped the event class of each data record very precisely. The preciseness of the result completely depends on the comparison analysis of the knowledge pattern of an event class with data records. The term sets constructed from each data records show a better comparison of term frequency association. Since, with 2000 set of test dataset the deviation in term of non-association is low and intern enhance the result precision and accuracy to high.

But, in case with a varied number of records to 10,000 the constructed term sets also increases and which shows a high variation in frequency association count in relation to the mapping class. This shows a low result in precision and accuracy with a variation of data records.

Comparison analysis PSA with Ensemble classifiers

Ensemble methods are statistical and computational learning procedures reminiscent of the human social learning behavior of seeking several opinions before making any crucial decision [33]. We utilize four major ensemble algorithms "AdaBoost (AB)", "Logic Boost (LB)" and "Voting (VOT)" to evaluate the improvisation of the proposed PSA method. We compare the accuracy and F-Measure of these ensemble methods with the PSA method to present the significance of the proposal.

Figure 9 shows the accuracy comparison between the ensemble classifier and PSA method-based classification. Here, PSA shows an average of 2% better classification accuracy in comparison ensemble classifiers. The generation of knowledge patterns for closely understanding the entities (records) relation with the event classes shows an enhancement in the classification accuracy. The accuracy of all ensemble methods shows similarity for initial data records but later with increasing data records the accuracy decreases. As all these methods utilize random sampling of features, so each of these might contains some duplicates of certain training points and leave out other training points completely. These reduce the accuracy of the ensemble classifiers in comparison to PSA.

Figure 10 shows the comparison of the F-measure result between ensemble classifiers and PSA-Method. As the classification accuracy of the PSA is much better than the ensemble classifiers, so the precision is also better compared to the ensemble classifiers. As each classifier's precision is completely depends on its trained knowledge. The PSA-trained knowledge provides a more precise mapping of the data records makes an improvement in the harmonic mean measure as F-measure. So, random sampling of features makes it different and uncorrelated leads lower on F-Measure in comparison to PSA.

Most of the ensemble classifiers are combines with multiple functions of the classifiers to enhance the accuracy of the prediction. In analysis of the different data records their accuracy of prediction turnout to be week and show low accuracy in compared to PSA classifier with increasing number of data

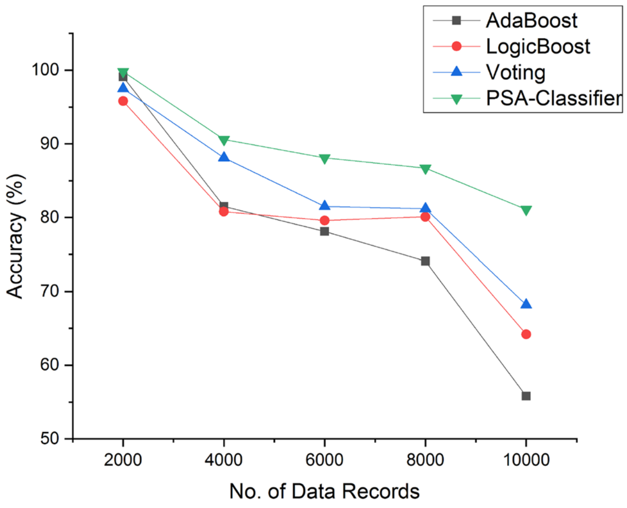


Fig. 9 Accuracy comparison between Ensemble classifiers and PSA

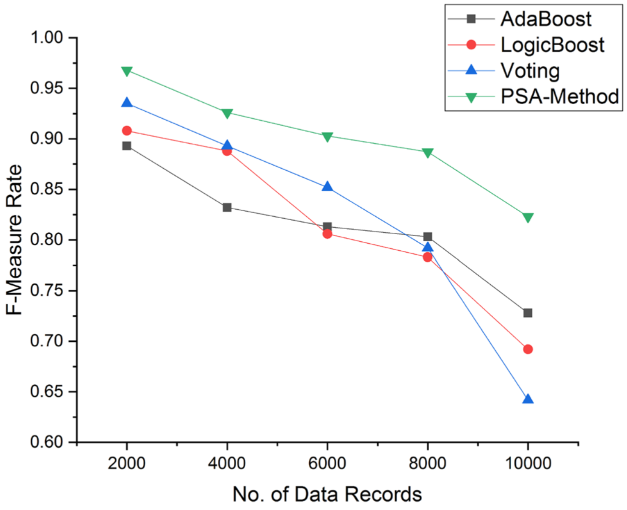


Fig. 10 F-Measure comparison between Ensemble classifiers and PSA

records. The PSA algorithm event class knowledge patterns derive it to better 2% better classification in the case of initial 2000 data records. In case of higher data records of 10,000 all the classifiers show a dip in the accuracy but PSA due to its top 3 highest frequency association count class prediction from the event class association and using Probabilistic Semantic Association to get the class having highest PSA association value for the data record allows it to achieve the better result from the ensemble classifiers.

To perform the state of art analysis we compare the Split and Combine Linear Discriminate Analysis (SC-LDA) algorithm with PSA. Seng et al. [35] introduce a novel big feature data analytics scheme for integration toward data analytics with the decision. In this scheme, a split and combine approach for a linear discriminate analysis (LDA) algorithm is termed SC-LDA. We measure the recognition accuracy to analyze the integration and classification efficiency of PSA in comparison to SC-LDA with a 0-split configuration.

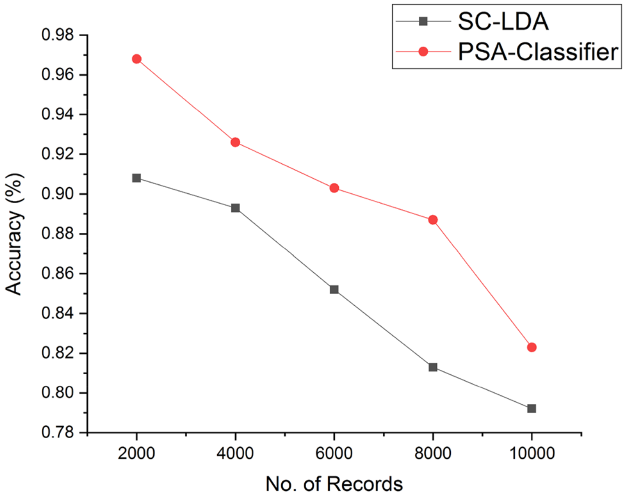


Fig. 11 Accuracy comparison between PSA and SC-LDA Algorithm

Figure 11 shows the accuracy performance for the PSA with SC-LDA Algorithm. The obtained result of the proposed PSA shows an average of 5% of improvisation in the prediction accuracy. The improvisation achieved by PSA is due to the semantic prediction accuracy with event categories, whereas SC-LDA shows lower due to its linear independence conditions. The prediction accuracy of PSA enhances the mapping of data records, which in turn improves the data integrations.

Intuitively, a probabilistic schema mapping describes a probability distribution of a set of possible schema mappings between a source schema and a target schema. The proposed PSA can answer the complexity of classification in terms of the size of the data complexity and the size of the mapping to the event classes. In comparison to the state-of-arts methods, it shows an average of 98.91% of mapping with 94.68% accuracy and 5.32% of error makes it much reasonable for integration in comparison.

The developments of the accuracy in terms of Principal Component Analysis (PCA) are due to the effectiveness of the pattern generation for different event classes. The generated pattern knowledge provides a broad range for relating the events entities with the pattern to have a clear and accurate mapping between events and event classes. The analysis with the huge collection of crime dataset using the proposed PSA builds the most significant knowledge which can be applied to incoming data class prediction for various prediction task to benefit for accurate prediction for heterogeneous big data sources which will even also supports for the appropriate data integration.

To illustrate the cost-effectiveness of the proposed PSA method we measure the time complexity using $O(n)$ notation in a linear search manner. Here, n is the number of elements in the event class array. Since it needs to compute certain parameters for every event mapping to its class. However, the number of such mapping is exponential in the size of the input dataset. It traverses the event class array sequentially to locate the required element. It searches for an element by comparing it with each element of the array one by one. In the best possible case, the element being searched may be found at the first position. In this case, the search terminates in success with just one comparison. Thus in the best case, the linear search algorithm takes *O(1)* operations. If the element being searched may be present at the last position or not present in the array at all. In the former case, the search terminates in success with n comparisons. In the latter case, the search terminates in failure with n comparisons. Thus in the worst case, the linear search algorithm takes *O(n)* operations.

The above result observations show an enhancement in the accuracy but it also shows the drop in accuracy with the increasing number of events records as a limitation. The reason for this was a complex and dynamic large and growing source linkage scheme. Sources are growing with time with diverse relationships, so it is not possible to change the pattern in the database frequently because of databases. So, it causes in to reduce the accuracy in classification and might affect the reliability and integrity of infrequent data changes.

Conclusion

A traditional data integration system is a system based on an advanced design that connects a limited number of sources and is relatively stable and usually takes hours. Using a probabilistic review approach allows potential conformances to be made and decision-making values to be decided. This paper offers an integrated and classification approach for developing and predicting the probability class over big data. This method explains the process of integrating and classifying big data through PSA. To generate a single data prediction, we supplement the traditional NB method with a semantic combination prediction. A large amount of SFPD data is evaluated to prove the effectiveness and improvisation of the proposal. The measured metric observation results show that compared with the NB classification an average increase of 10%, and an average increase of 12% with F-measure improvisation. In the future, the proposed method can be applied for accurate data mapping for the various prediction tasks to benefit from the accurate prediction of heterogeneous large data sources.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Lcuadrado JL, Carrasco IG, Hernández JLL, Fernández PM, Fernández JLM (2020) Automatic learning framework for pharmaceutical record matching. IEEE Access 8:171754
2. Yan G, Wang H (2020) Autonomous coordinated control strategy for complex process of traffic information physical fusion system based on big data. IEEE Access 8:148370
3. Mao R, Xu H, Wu W, Li J, Li Y, Lu M (2015) Overcoming the challenge of variety: Big data abstraction, the next evolution of data management for AAL communication systems. IEEE Communication Magazine 53(1):42–47
4. L. da Silva Daniel, L. P. Pedro, S. L. Stanzani, A. Paulo, A. C. Sheffer (2014) A computational framework for integrating and retrieving biodiversity data on a large scale. IEEE International Congress on Big Data
5. Adnan K, Akbar R (2019) An analytical study of information extraction from unstructured and multidimensional big data. Journal of Big Data 6(1):91
6. Dong XL, Srivastava D (2015) Big data integration. Morgan & Claypool 1:1–98
7. Gu B, Li Z, Zhang X, Liu A, Liu G, Zheng K, Zhao L, Zhou X (2017) The Interaction between schema matching and record matching in data integration. IEEE Trans Knowl Data Eng 29(1):186–199
8. Y. Roh, G. Heo, S. E. Whang (2019) A survey on data collection for machine learning: a big data integration perspective. IEEE Transactions on Knowledge and Data Engineering, 1–1
9. Paniagua VS, Zavala RMR, Segura-Bedmar I, Martínez P (2019) A two-stage deep learning approach for extracting entities and relationships from medical texts. Journal Biomed. Information 99:103285
10. B. Louie, L. Detwiler, N. N. Dalvi, R. Shaker, P. Tarczy-Hornoch, D. Suciu, (2007) Incorporating uncertainty metrics into a general-purpose data integration system. 19th International Conference on Scientific and Statistical Database Management (SSDBM) 19–19
11. Desmet B, Hoste V (2018) Online suicide prevention through optimized text classification. Information Science 439–440:61–78
12. Gong P, Cao Y, Cai B, Li K (2018) Multi-information location data fusion system of railway signal based on cloud computing. Future Gener Computer System 88:594–598
13. B. Marthi, B. Milch, S. Russell (2003) First-order probabilistic models for information extraction. In IJCAI workshop on learning statistical models from relational data
14. Califf ME, Mooney RJ (2003) Bottom-up relational learning of pattern matching rules for information extraction. J Mach Learn Res 4:177–210
15. Sun Y, Bie R, Zhang J (2016) Measuring semantic-based structural similarity in multi-relational networks. International Journal of Data Warehouse and Mining 12(1):20–33

16. Zhang Y, Wu H, Sorathia V, Prasanna VK (2013) Event recommendation in social networks with linked data enablement. In ICEIS Conference 2:405
17. Shvaiko P, Euzenat J (2005) A survey of schema-based matching approaches. Springer J Data Semantics. https://doi.org/10.1007/11603412_5
18. S. Bergamaschi, L. Po, S. Sorrentino (2007) Automatic annotation in data integration systems. OTM Workshops Springer Berlin Heidelberg Berlin, Heidelberg pp 27-28
19. M. Magnani, N. Rizopoulos, P. McBrien, D. Montesi (2005) Schema integration based on uncertain semantic mappings, In Proc. of Conference on Conceptual Modeling, Springer Berlin Heidelberg, Berlin, Heidelberg
20. Doan A, Halevy A, Ives Z (2012) Principles of Data Integration. Morgan Kaufmann, Waltham, MA, USA
21. A. Jinchuan, C. Yueguo, Xiaoyong, Li. Cuiping, L. Jiaheng, Z. Suyun, Z. Xuan (2013) Big data challenge: a data management perspective. Front Computer Science*******
22. Sun Y, Bie R, Zhang J (2016) Semantic relation computing theory and its application. In J Net Comput Applicat 59:219–229
23. Sun Y, Jara AJ (2014) An extensible and active semantic model of information organizing for the internet of things. Pers Ubiquit Computing 18(8):1821–1833
24. M. D. Lee, B. Pincombe, and M. Welsh (2005) an empirical evaluation of models of text document similarity. Mahwah, NJ: Erlbaum, pp. 1254–1259
25. Zhang J, Yao C, Sun Y, Fang Z (2016) Building text-based temporally linked event network for scientific big data analytics. Pers Ubiquit Comput. https://doi.org/10.1007/s00779-016-0940
26. M. A. Hasan, V. Chaoji, S. Salem, M. Zaki (2006) Link prediction using supervised learning. In Proceedings of SDM-06 workshop on Link Analysis, Counterterrorism and Security
27. E. Agichtein, V. Ganti (2004) Mining reference tables for automatic text segmentation". In Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 20–29
28. G. Kumaran and J. Allan (2004) Text classification and named entities for new event detection, In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 297–304
29. N. N. Dalvi, D. Suciu, (2007) Management of probabilistic data: foundations and challenges", PODS, pp. 1–12, ACM
30. Bishop CM (2006) Pattern Recognition and Machine Learning. Springer, New York, NY, USA
31. SFPD Datasets: "City and County of San Francisco-SF Open Data". https://www.kaggle.com/c/sf-crime/data, May 2015.
32. N. Rizopoulos, P. McBrien (2005) A general approach to the generation of conceptual model transformations, In Proc. CAiSE. LNCS
33. Zhang C, Ma Y (2012) "Ensemble Machine Learning: Methods and Applications", Springer: Boston. MA, USA
34. Bovenzi G, Aceto G, Ciuonzo D, Persico V, Pescapé A (2020) A big data-enabled hierarchical framework for traffic classification. IEEE Trans Netw Sci Eng 7(4):2608–2619
35. Seng JKP, Ang KL-M (2017) Big feature data analytics: split and combine linear discriminant analysis (SC-LDA) for integration towards decision making analytics. IEEE Access 5:1405
36. Liu W, Wu W, Wang Y, Fu Y, Lin Y (2019) Selective ensemble learning method for belief-rule-base classification system based on PAES. Big Data Mining and Analytics 2(4):306–3018
37. Zhang C, Cui C, Gao S, Nie X, Xu W, Yang L, Xi X, Yin Y (2019) Multi-Gram CNN-based self-attention model for relation classification. IEEE Access 7:5343–5357
38. Reddy GT, Reddy MPK, Lakshmanna K, Kaluri R, Rajput DS, Srivastava G, Baker T (2020) Analysis of dimensionality reduction techniques on big data. In IEEE Access 8:54776–54788
39. N. Deepa, Q. -V. Pham, D. C. Nguyen, S. Bhattacharya, B. Prabadevi, T. R. Gadekallu, P. K. R. Maddikunta, F, Fang, P. N. Pathirana (2021) A survey on blockchain for big data: Approaches, opportunities, and future directions. In Cryptography and Security
40. Tang M, Alazab M, Luo Y (2019) Big data for cybersecurity: vulnerability disclosure trends and dependencies. IEEE Trans Big Data 5(3):317–329
41. Khan RU, Zhang X, Kumar R, Sharif A, Golilarz NA, Alazab M (2019) An adaptive multi-layer botnet detection technique using machine learning classifiers. Appl Sci 9(11):1–22
42. Gadekallu T, Rajput D, Reddy P, Lakshman K, Bhattacharya S, Singh S, Jolfaei A, Alazab M (2020) A novel PCA–whale optimization-based deep neural network model for classification of tomato plant diseases using GPU. J Real-Time Image Proc 18:1–14