International Journal of Engineering & Technology, 7 (4) (2018) 4429-4434

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 17144
Research paper



A classification model on probabilistic semantic relation for big data: an integrated approach

Madhu M. Nashipudimath [1] *, Subhash K. Shinde [2]

[1] Research Scholar, Faculty of Computer Engineering, Pacific Academy of Higher Education and Research University, Udaipur, India
[2] Lokmanya Tilak College of Engineering, Navi Mumbai, India
**Corresponding author E-mail: madhu.nashipudi@yahoo.in*

Abstract

Data mining is process of analyzing information repositories. As data store took shape of big data, it is difficult to find relevant patterns with current techniques. Existing framework don't suit integration and analysis of complex scenario. This insufficiency motivates to propose new solutions. The major problem with big data integration and analysis is due to complex interdependence between the changing data granularity, incompatible data models, and data contents. Hence integration and classification model based on probabilistic semantic relation (PSR) of attribute pattern for big data source is proposed. It learns interrelationships and interdependence pattern among data class and data source. This knowledge is utilized to classify probabilistic relation prediction among the pattern and source data which helps in data classification and future analysis. The model implements Data integration and mapping, Construction of knowledge base, and Naive based (NB) PSR approach. An experiment is done over real crime dataset. Measures like Precision, Recall, Fall-out rate and F-measure are calculated to evaluate results. Experiment shows average of 10% increase in precision and recalls as compared to NB classification and an average of 7% improvisation in F-measure. This improvisation suggest that proposed model can be applied to future data class prediction for various prediction task.

Keywords: *Big Data; Integration; Probabilistic Relation Prediction; Semantic Classification.*

1. Introduction

This Big data is ubiquitous and being utilized in many areas, which includes computer vision, machine learning, and financial analysis etc. Social network is one of the main sources among these to facilitate individuals to interact with web. It als o helps and promotes community creation, sharing ideas, opinions, and contents. Social networks had revolved out to be the medium of numerous applications such as targeted advertising and referral service, collaboration filtering, Behaviour modelling, forecasting, analysis and identification of aggressive behaviour, cultural trend monitoring, epidemic research, and reading and tracking crowd moods, even the revelation of terrorist networks activities and political deliberations [1-2].

The existing big data frameworks are concerned about very large and complex database collection of the Facebook; Twitter, and LinkedIn. Monthly, almost 1 billion users provide tweets and links throughout the world [3-4]. This large data set information must be collected, stored, transmitted, retrieved, shared, analysed and processed in an efficient manner. Collecting and integrating immense amounts of data [5] to extract information and knowledge is not possible using traditional methods [6] and tools [7]. Information exchange system and a unified interface [8] to multiple sources is required with the capability to distribute information in multiple systems [9]. Currently, data integration schemes [10] are inherently a natural addition to existing database systems, where uncertainty are specified in a structured format and the data is represented in one of the existing data models, such as relational or Extensible Markup Language (XML). The data integration system also has accurate information regarding the mapping data source to the arbitration schema through appropriate data integration systems [10].

Often the information collected is not formatted or ready for analysis. Hence it is essential to keep track of the progress of an incident whenever it occurs. In general, there are strict time limits on the demands of tracking events. In particular, the incident is an adversity or an emergency. Hence it is essential to identify the incident and time information quickly to manage the progress of an accident, including avoiding disasters or improving results [11-13]. In case of emergencies, enough data about the emergency should be gathered as soon as realizable and then analyze the information to sustain the decision. Historically similar incidents can be used to predict and guide the occurrence of current incidents. Unprocessed data with raw format cannot be used to analyze effectively. Rather, it needs an information extraction process that extracts the necessary information from the underlying source and presents it in a structured format suitable for analysis.

Big data-oriented information is necessary to combine both structured and unstructured data for analysis. Structured data with specific qualities and values [8] can be used directly in the information analysis and it is based on calculations for quantitative analysis [14]. Extracting unstructured data from text and change it to structured data is through with mostly using natural language processing techniques. Structured data shall contain clear attributes and values, while unstructured data should be organized, extracted and transformed into structured data for information analysis.

Present effort is to integrate heterogeneous data from different data sources and classify it based on probabilistic semantic rela-

Copyright © 2018 Madhu M. Nashipudimath, Subhash K. Shinde. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

tionships (PSR) analysis. It can predict and guide the occurrence of current incidents.

Probabilistic Semantic Relation for Integration and Classification over Big Data is proposed for data integration and classification. The data integration systems will extend the architecture of existing data integration systems of Hadoop Framework to build the required modules and data schema. The classification method extends a Naive based probabilistic approach utilizing trained dataset patterns, which allows annotating the anonymous data automatically for data sources. The model implements three methods to meet the objective, (1) Data integration and mapping, (2) Construction of knowledge base, and (3) Naive based probabilistic semantic relation (NB-PSR) approach.

The background study with respect to big data integration and semantic relation is discussed in Section II, Section-III explains the proposed integration and classification model, Section-IV investigates the experiment evaluation and Section-V presents the conclusion of the paper.

2. Background study

Data management and analysis of big data experienced some problems. It was specifically in managing the traditional form of relational databases for the multi-source and heterogeneous information. The reason was the large and scalable source relational schema which is complex and dynamic. Resources may progress with time and on their relationships. It is impossible to change the schema of the database frequently because of relations between the databases and the application is based on schema. Relational databases are somewhat appropriate for maintaining reliability and integrity throughout frequent data alterations. But it may not suitable for information analysis that believes more relationships involving different entities than data preservation.

J Chen et al. [22] presents the reviews on "Big data challenge: a data management perspective" in particular to big data diversity, reduction, integration and cleaning, indexing and query, and finally big data analysis and mining. It mainly focuses on the challenge of the traditional machine learning and statistics algorithms on big data. But, it shows a limitation to present the corresponding approaches and mechanism comparison in related to data integration and indexing in big data.

In order to manage data for information analysis, one promising approach is to use hierarchical models (such as tree models) and network models to construct multi-source and heterogeneous information. Hierarchical models and network models can characterize complex relationships between entities. However, the meaning of a fragment composed of entities and relationships is required for domain-specific applications. It is needed to study a new data model or information organization model for managing and analyzing large amounts of data.

M.Wang et al. [5] proposed a similarity join techniques for big data integration and an estimate-based algorithm for load balance. A combines multiple filter algorithms and uses MapReduce to process similarity join tasks in parallel enhance the efficiency of the Integration. A combines multiple filter algorithms and uses MapReduce to process similarity join tasks in parallel enhance the efficiency of the Integration. But it not explore the other similarity functions and conditions which can provide to the lower threshold and large documents for the MapReduce framework.

Some of the semantic data models [3] for heterogeneous data sources and the characteristics of the times are being proposed to meet the big data. X. L. Dong et al. [6] presents the challenges of big data integration such as schema mapping, record linkage and data fusion. It describes the challenges using examples and techniques for data integration for addressing the new challenges raised by big data. It clearly states the problems to integrate the existing data and systems, but it does not make clear any real time based data source analysis and integration of big data.

Analyzing information based on semantically organized information or patterns can facilitate in finding important information for decision making [15]. Here author presents an experiment based on semantic similarity measure to address the problem of the decision support systems dealing with text documents. It utilized a Latent Semantic Analysis (LSA) approach to correlated the corpus of the document. The experimental result shows the best performance to the document having well defined structured and related document. But this system not evaluated for any big document collection. A need of search engines or any text based decision systems keeps this approach open for further exploring.

Initially it is imperative to explain the difficulties caused by the restrictions of storage and computing in order to utilize large amount of data. Because one issue is that big data is scalable and on the other side, the information required for the analysis must respond quickly. Fortunately, cloud computing technology enables to process analysis and evaluation [16] vast amounts of information to discover knowledge and intelligence [17].

In order to use large data, this work is concerned in events or knowledge [29] implicit in big data rather than huge data details. An Incident is a semantic unit to understand contexts that include elements that formulate events such as when, where, who, why, how and what and hidden knowledge in large data. It can guide to the most important value in the characteristics of big data [18]. Every day and every moment, one or more events occur in the world which will remain in the history and memory of the human form. Many events are closely related to our daily living and work, and some of the major events make immense impact on us in the future prospect.

The semantic of the event is strongly related to the language as it is one of the most significant ways to correspond to the meaning. So the event semantic has become a branch of the assumption of semantic. The origin of its proposal is that Davidson verbs special " hidden " place for the argument [19, 20]. This suggestion has proved very productive in the natural language with an extensive variety of significances. Also, the message expressed by the language difference is extracted automatically. Events and expression of semantics of such events are carried by the processing of natural language understanding [21].

The event plays a significant responsibility in data analysis [13]. In order to understand the situation and make appropriate decisions, it is essential to clarify the progression of events [22] and organize relevant information resources according to the temporal order of events. During the incident, the people involved will do something somewhere. Many events have formed part of history together. Therefore, the information analyst urgently needs information resources composed of other events organized. Incident-Based Information Organizations are naturally sequential clues to categorizing information in the circumstance of incidents. However, incidents do not have a uniform definition. For example, in a language study, an event can be a verb or a noun. In industrial control, an incident can be a transform of the state. Though the definition of an event is different, every event is inseparably related to two indispensable elements or attributes, time and place. It is ubiquitous to have an inquiry and study of the temporal and geographic information of incidents. Time is also an integral part of every information space.

Incidents are similar to the semantic association of information resources [23], and there have a different category in different intelligent applications. Incident and time information is often used for information retrieval and question answering. Large-scale data of online human interaction has been challenging to study interdisciplinary theories that describe the intrinsic co-operation that is multidimensional [24].

The proposed model is based on event-based dataset analysis, discussed in detail in the next section. It also leverage to extend the existing machine learning and data mining techniques [21] to facilitate rich analysis of integrated data sets.

3. Proposed integrated and classification model

The proposed model consists of two phase of functionality as, Integration of big data sources to structured data, and classification of data for analysis as shown in Fig.1. In the first phase, it performs the integration of big data sources to structured data through implementing the pre-processing mechanism for different data source which will be used for learning and analysis. The classification and prediction of structured data are performed for analysis in the second phase.

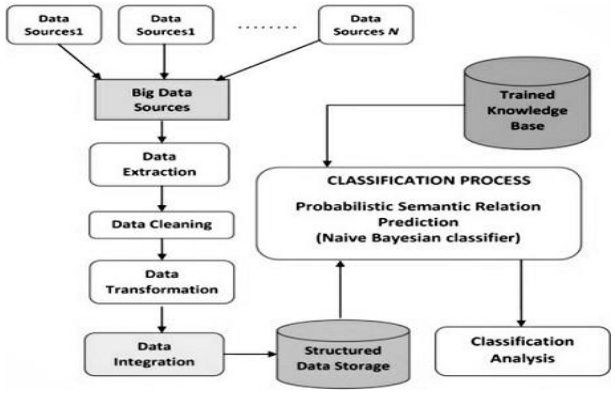


Fig. 1: Integration and Classification Model.

The model is evaluated over big datasets published by SFPD[28]. The dataset is captured through a Crime Incident Reporting System over the different location. It predicts the incidents based on the activist information, which will be important for the monitoring and analysis of incidents. To facilitate this prediction analysis over the datasets, a probabilistic semantic relation interpretation is performed among the captured incident description to predict the category of incident. In the following section, the methodology of Data integration and mapping, Construction of the knowledge base and Naive based probabilistic semantic relation (NB-PSR) approach are discussed in detail.

3.1. Data integration and mapping

The big data is a source of uncertain data recorded from different stream and media. The recorded data may have much uninterested data which need to filter and cleaned before performing the analysis. Data integration mechanism performs "data extraction" and "data cleaning" methods to construct the structured data. Mostly, the information collected is not in ready for analysis. It will be likely to be incomplete, noisy, and unpredictable. Data cleaning or cleansing schedules effort to load in missing values and inconsistent information. Cleaning of data schedules tries to fill up missing values, smooth out noise while categorizing outliers, data, and accurate data predictability.

3.1.1. Data extraction

This method which typically integrates the different data source into a required unified structure. It also updates or loads the regular data from the sources. The extraction of data generally performs using an application interface. The model uses Hadoop Query Editor Interface to extract and load data to a designed database schema.

3.1.2. Data cleaning

The process of data cleaning fills the missing values for the attributes and removes the irrelevant tuples. It detects errors or duplicate tuples in the data and resolves them wherever is possible. The resulted cleaned data will be effectively used for integration. Most mining routines have procedures for handling incomplete or noisy data, but these are not always powerful. As a result, useful pre-processing steps are executed over data through the data clean-up routine.

This work includes data from multiple sources in form of data files. Even though the attributes presented in the file might be same but the values it stores cause inconsistent and redundancy. It replaces all missing category attribute values by the same constant value as “None”.

3.1.3. Data transformation

The data transformation process transforms or consolidates the data into appropriate forms. A smoothing process is applied to specify transformations to correct data inconsistencies. For example, a crime category "Vehicle Theft" in one data file is "Vehicle Thft" and in another it may be "Veh. Theft" or "Vh. Theft ". These data field is converted to a unique value to avoid the redundancy of categorization. It also applied for the various unstructured crime posts description, as for the same crime category different descriptions are posted.

3.1.4. Data integration

In data integration, it merges the cleaned data commencing from numerous sources into a consistent data accumulation. The data sources incorporate in this model is the multiple CSV files. The data are stored in an appropriately structured form for analysis. A multiple crime data files are obtained from different location to successfully integrate into a coherent data store in this work.

The SFPD Data [28] contains 13 years crime data information year wise of different location. The integration method is applied over this datasets to integrate this 13 years data to commence proper analysis as depicted in Fig 2.

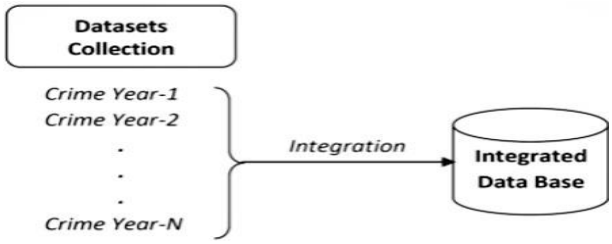


Fig. 2: Integration Model of Dataset.

3.1.5. Data mapping

A set of data category classes is built for correlation and classification. Each data record can be mapped to a predicted category class. The described attributes are mapped to respective category class through a training process through training process.

3.2. Construction of pattern knowledge base

The construction pattern knowledge base is performed over transformed data storage [25-26]. It utilized 40% of the historical data to build the pattern knowledge base for each individual incident category class, while the rest and latest data of incidents can be classified for prediction using the pattern knowledge base. It is the source of information analytics and decision making. The training construction module performs an individual data training to construct a pattern which can be used for real-time streaming data. The knowledge patterns are utilized to show the estimate the fascinating of resulting patterns.

This process is a supervised learning process [26] as the class label of everyone training [20] tuples is provided. It uncovers appealing data patterns unknown in large training data sets. For each category, class label patterns represent a useful knowledge. It performs a frequent item set mining to construct the associated values in the trained dataset as described in Algorithm-1.

Algorithm-1: Construction of Knowledge Base

Input: Trained Datasets, T, and Crime Category, C.

Output: $P_i \rightarrow$ Knowledge Patterns.

D → collection of trained data tuples from category c_i
S → collection of split terms of d_j
Ci → category value of each Crime Category from C.
Dj → each data tuple from D.
SK → each term value from S.
For i=0; each category c_i in C; i++;
{
D ← read the tuple from T where category is c_i
For j=0; each tuple d_j in D; j++;
{
S = split (d_j, delim);
For k=0; each term s_k in S; k++;
{
Fcnt = 0;
If SK ϵ D then
fcnt++;
End if
If fcnt > 1 then
P_i [i] = {SK}
End if
}
}
}
Return P_i;

The constructed knowledge pattern, P_i, will be the basis of prediction on the current incident, while the data of the concluding kind are the indications of analysis on the present incident. For the classification and prediction, a modified Naive Bayesian algorithm [30] is suggested with the integration of semantic relation probability [24] among knowledge pattern[27] and the test data.

3.3. Naive based probabilistic semantic relation (NB-PSR) approach

The process of classification usually generates lot of probability uncertainty for class prediction. This model proposes a Naive based probabilistic semantic relation (NB-PSR) approach for automating the incident prediction. It computes a probability association preference based semantic mappings that specify the relationship between the constructed knowledge base and the current data.

Algorithm-2: NB-PSR Approach
Input: Z, P, C and Trained Datasets, T.
Output: d_{nc} → predicted Category Class.
Z → Collection of current data.
P → Generated Knowledge Patterns.
C → Collection of Crime Category.
d_n → each data record from Z.
p_k → each category pattern from P.
t_i → each term from the split data record d_n.
For n=0; each current data d_n ϵ Z; n++;
{
For k=0; each category pattern p_k ϵ P; k++;
{
D_T = split (d_n, delim);
asso_cnt=0;
For i=0; each term t_i ϵ D_T; i++;
{
If t_i ϵ p_k then
asso_cnt++;
}
// -- category association count--
d_cat_assoc[n] = asso_cnt;
}
//-- Get top 3 highest associated category from C --
H [] = getTop (3, d_cat_assoc[ ], C);
//-- Probabilistic Semantic Relation --
psr_tot=0;
For j=0; each assoc_category a_i ϵ H []; j++;
{
D_T = split (d_n, delim);
psr_val=0;
For i=0; each term t_i ϵ D_T; i++;
{
psr_val = Σ t_i ϵ T;
psr_tot=psr_tot+psr_val;
}
psr_assoc_cat [j] = psr _tot;
}
d_{nc} = getMax_PSR_assoc (psr_assoc_cat [ ], H[ ]);
}

From each probabilistic semantic relation calculated for the data record of Z is used for different analysis purpose. The model performs an integration and classification analysis over the SFPD crime information datasets as shown in Fig.3. to evaluate the proposal.

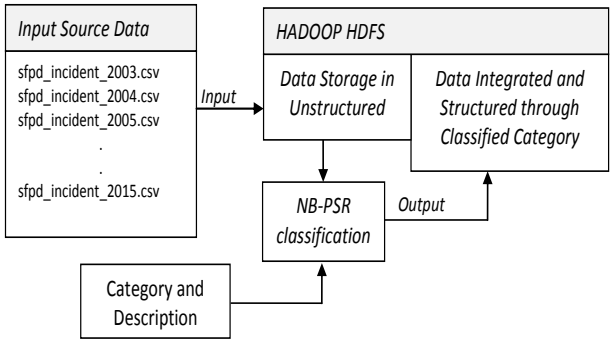


Fig. 3: Input and Output Details of NB-PSR Classification.

4. Experimental evaluation

4.1. Data sources

Dataset is collected from SFPD openData[28] to evaluate the proposed model. The dataset is a collection of crime incidents from the year 2003 to 2015 across 10 districts of San Francisco. The collection provides a total of 15,86,225 crime incidents and 11 data fields for each incident. The work mainly focuses on two data fields for classification, which provide the incidents category and its description.

Category: It provides the incident category class identified. It has 36 crime categories which are being extracted from datasets.

Description: It provides the incident information narrative. A sample of data is shown in Table-1.

Table 1: Category and Description

Category	Description
Vehicle Theft	Stolen Automobile
Warrants	Enroute To Outside Jurisdiction
Other Offenses	Traffic Violation Arrest
Warrants	Warrant Arrest
Secondary Codes	Domestic Violence
Assault	Threats Against Life
Assault	Aggravated Assault With A Deadly Weapon
Suspicious Occ	Investigative Detention
Larceny/Theft	Grand Theft From Unlocked Auto
Drunkenness	Under Influence Of Alcohol in a Public Place

4.2. Implementation and evaluation measures

Datasets are extracted and collected manually[28] for evaluation. Data cleaning is carried out to remove data errors and noisy data. The clean data is uploaded to a default schema of Hadoop Framework. A sfpd_db database schema is created to migrate the data from default schema to the sfpd_db database. The migration of data is performed through a SQL script.

A Java based program is implemented to process the data and generate each category knowledge pattern. The knowledge patterns are stored in relation database schema. Testing is carried out by the traditional Naive Bayesian (NB) and proposed NB-PSR

classification method over a different range of data records. This is used to measure the effectiveness of the proposal. The most popular measures to evaluate the classifier efficiency are Precision, Recall, Fall-out rate and F-measure. To evaluate the outcome results of the dataset these measures are performed. The measures computed based on the confusion matrix values are identified during the process. The following observations as shown in Table 2 are considered for the calculation.

Table 2: Category and Descriptionn

True Positive (TP)	No. of data record correctly predicted.
True Negative (TN)	No. of data records Incorrectly predicted.
False Negative (FN)	No. of data records Not predicated.

Based on the observation following are computed,

$$\Pr ecision = Tp / Tp + TN \quad (1)$$

$$\mathrm{Re}\, call\,(R) = TP / TP + TN + FN \quad (2)$$

$$FallOut\ _\ Rate = TN / TP + TN \quad (3)$$

$$F - Measure = 2 \times (P \times R) / (P \oplus R) \quad (4)$$

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

4.3. Evaluation results

Model performs the classification and prediction comparison of unclassified datasets class utilizing the traditional NB and NB-PSR to measure the improvisation. NB classification is mostly being used in text classification and prediction tasks. The prediction comparisons of results for correct, incorrect and not predicted are presented in Fig.4. It shows that NB-PSR has improvisation over NB incorrect prediction.

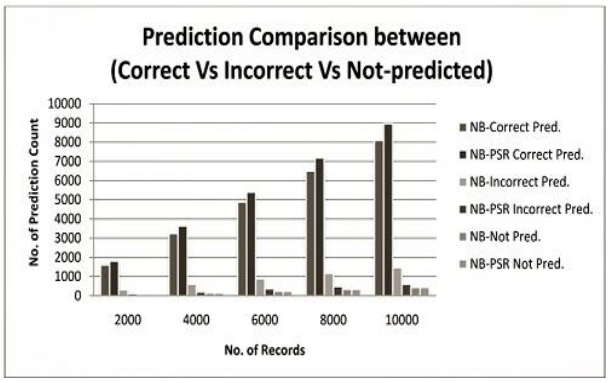


Fig. 4: Prediction Comparison between NB and NB-PSR.

Precision and recall rate are calculated to measure positivity and sensitivity. Fig.5. and Fig.7. represent the precision and recall rate comparison. The proposed NB-PSR shows an average of 10% improvisation in precision rate and recall rate as compared to NB. The improvisation is due to the enhancement in computation through probabilistic semantic relation.

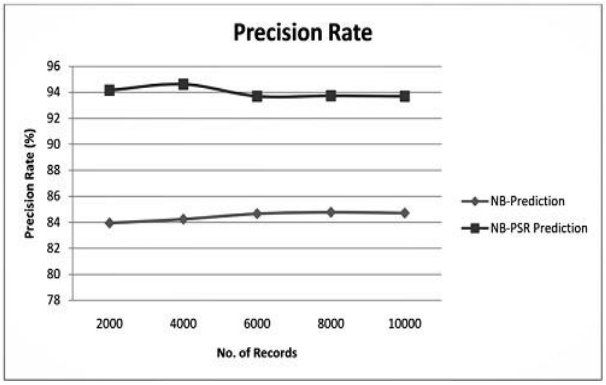


Fig. 5: Precision Rate Comparison between NB and NB-PSR.

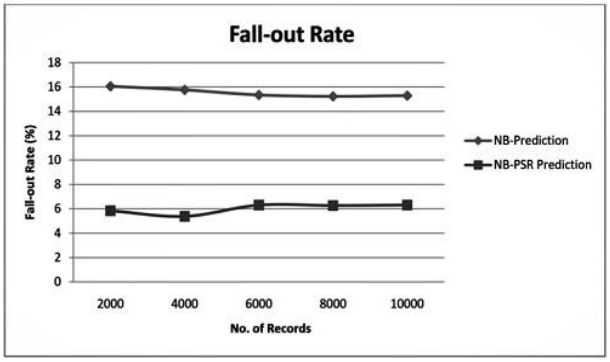


Fig. 6: Fall Rate Comparison between NB and NB-PSR.

Fig.6. and Fig.8. present the fall-out and F-measure rate. Fall-out rate measures the proportion of incorrect prediction against the overall data record predicted.

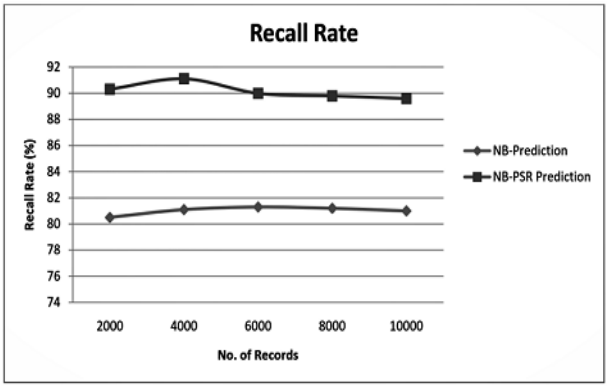


Fig. 7: Recall Rate Comparison between NB and NB-PSR

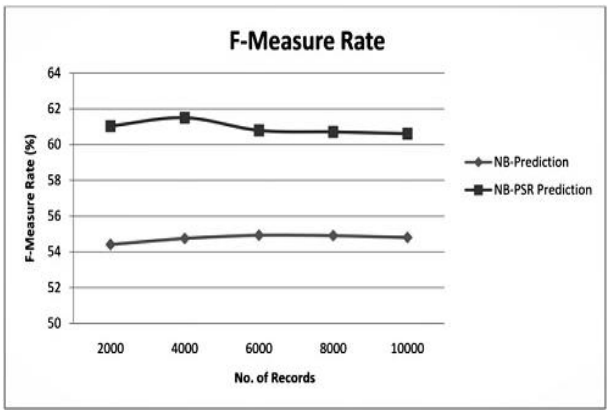


Fig. 8: F-Measure Comparison between NB and NB-PSR.

NB-PSR shows an average of 9% low fall-out rate in comparing to NB. In the case of F-measure which combine precision and recall to measure the harmonic means shows an improvisation in compared to NB classification.

All the above result observation proves the effectiveness of data integration and it applicability of the proposal for the feature analysis and decision-making tasks.

5. Conclusion

Usually a traditional data integration system is a scheme that interconnects a limited number of resources. It is built on a conventional design that is relatively stable and typically takes hours. To overcome this problem, a new model is proposed. In this paper an integrated and classification model is implemented for processing and predicting the probability class over big data. The model explains the process of big data integration and probabilistic semantic relation classification. Utilizing a probabilistic analysis approach allows to perform prospective equivalent and allocate probability values for decision making.

The traditional Naive Bayes (NB) method is enhanced with semantic relation prediction to improvise the integrated data prediction, which helps to predict and classify heterogeneous data. The experiment is carried out with standard dataset SFPD [28]. The measured metric observation shows an average of 10% increase in precision in compare NB classification and average of 12% improvisation in F-measure. By looking at classification performance, it is not surprising that the PSR produces better results than traditional Naive Bayes method. The proposed model can be applied to future data prediction for various prediction tasks. It can be effective and useful for prediction heterogeneous big data .

Acknowledgement

Madhu M Nashipudimath would like to thank Principal and Management of Pillai College of Engineering, New Panvel, India for their continue support and cooperation during this research work.

References

[1] Sun Y, Lu C, Bie R, Zhang J. Semantic relation computing theory and its application. In: Journal of Networking and Comput Application. 2016; 59:219-229. https://doi.org/10.1016/j.jnca.2014.09.017.

[2] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and S. Gerd. Evaluating similarity measures for the emergent semantics of social tagging. ACM in Proceedings of the 18th international conference on World Wide Web. New York, 2009; p.641-650.

[3] R. Mao , H. Xu , W. Wu , J. Li , Y. Li , and M. Lu .Overcoming the challenge of variety: Big data abstraction, the next evolution of data management for AAL communication systems. IEEE Communication. Magazine. Vol 53. No one.2015; p. 42 - 47.

[4] Nagwani, N.K.Summarizing large text collection using topic modeling and clustering based on MapReduce framework. In:Journal of Big Data. 2015; 2(1).p. 6. https://doi.org/10.1186/s40537-015-0020-5.

[5] Wang, M., Nie, T., Shen, D., Kou, Y. and Yu, G., November. Intelligent similarity joins for big data integration. In: 10th Web Information System and Application Conference (WISA). IEEE 2013. p. 383-388.

[6] X.L. Dong, D. Srivastava. Big data integration. In: IEEE International Conference in Data Engineering (ICDE). 2013. 29:1245-1248. https://doi.org/10.1109/ICDE.2013.6544914.

[7] Daniel L. da Silva, Pedro L. P.,Silvio L. Stanzani, Paulo A, A. Sheffer C. A Computational Framework for Integrating and Retrieving Biodiversity Data on a Large Scale.In: IEEE International Congress on Big Data. 2014.

[8] Gu, B., Li, Z., Zhang, X., Liu, A., Liu, G., Zheng, K., Zhao, L. and Zhou, X., The Interaction Between Schema Matching and Record Matching in Data Integration. In: IEEE Transactions on Knowledge and Data Engineering, 2017 29(1): p.186-199. https://doi.org/10.1109/TKDE.2016.2611577.

[9] S. Bergamaschi, L. Po,S. Sorrentino. Automatic annotation in data integration systems.In: OTM Workshops. LNCS 4805, Springer. 2007. p. 27-B. Louie ,L. Detwiler ,N. N. Dalvi ,R. Shaker , P. Tarczy-Hornoch, D. Suciu. Incorporating uncertainty metrics into a general-purpose data integration system. SSDBM, 19, IEEE Computer Society, 2007. https://doi.org/10.1007/978-3-540-76888-3_14.

[10] Zhang, J., Yao, C., Sun, Y. and Fang, Z. Building text-based temporally linked event network for scientific big data analytics. Personal and Ubiquitous Computing. 2016 20(5):743-755. https://doi.org/10.1007/s00779-016-0940-x.

[11] W. Zhang, J. Wang, and Wei Feng. Combining latent factor model with location features for event-based group recommendation. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013. p 910-918. https://doi.org/10.1145/2487575.2487646.

[12] Yinuo Zhang, Hao Wu, Vikram Sorathia, and Viktor K. Prasanna. Event recommendation in social networks with linked data enablement. In: ICEIS Conference 2013.

[13] Sun Y, Bie R, Zhang J. Measuring semantic-based structural similarity in multi-relational networks. International Journal of Data Warehouse and Mining.2016; 12(1): p. 20-33. https://doi.org/10.4018/IJDWM.2016010102.

[14] Sun Y, Jara AJ. An extensible and active semantic model of information organizing for the internet of things. Personal and Ubiquitous Computing. 2014 18(8):1821-33. https://doi.org/10.1007/s00779-014-0786-z.

[15] Micheal D. Lee. Brandon Pincombe, Matthew Welsh. An Emperical evaluation of models of text document similarity. Inproceddings of the 27th annual conference of the Cognitive Science Society, 2005, pp. 1254-1259.

[16] Shvaiko, Pavel, and Jérôme Euzenat. A survey of schema-based matching approaches. In: Journal on data semantics IV. Springer Berlin Heidelberg, 2005. P 146-171. https://doi.org/10.1007/11603412_5.

[17] Magnani M, Rizopoulos N, Brien PM, Montesi D. Schema integration based on uncertain semantic mappings. In: International Conference on Conceptual Modeling 2005. pp. 31-46. Springer Berlin Heidelberg.

[18] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In: Proceedings of SDM-06 workshop on Link Analysis, Counter terrorism and Security. 2006.

[19] Popescul A, Ungar LH. Statistical relational learning for link prediction. In: IJCAI workshop on learning statistical models from relational data 2003 2003.

[20] Agichtein E, Ganti V.Mining reference tables for automatic text segmentation. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining. ACM. 2004. P 20-29. https://doi.org/10.1145/1014052.1014058.

[21] G. Kumaran and J. Allan.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, NY. 2004. p. 297-304.

[22] Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S. and Zhou, X., 2013. Big data challenge: a data management perspective. Frontiers of Computer Science, 2013; 7(2):157-164. https://doi.org/10.1007/s11704-013-3903-7.

[23] Dalvi, N. and Suciu, D. Management of probabilistic data: foundations and challenges. In: Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. 2007. p.1-12. https://doi.org/10.1145/1265530.1265531.

[24] Mary Alaine Califf, Raymond j. Mooney. Bottom-up relational learning of pattern matching rules for information extraction. In:Journal of Machine Leraning:2003:4:117:210.

[25] Califf M. E, Mooney R J. Bottom-up relational learning of pattern matching rules for information extraction. In: Journal of Machine Learning Research. 2003; 4:177-210.

[26] Marthi B, Milch B, Russell S. First-order probabilistic models for information extraction. In: IJCAI workshop on learning statistical models from relational data. 2003.

[27] SFPD Datasets: City and Country of San Francisco-SF Open Data. https://data.sfgov.org, Accessed May 2015.