Knowledge Fusion Technique Using Classifier Ensemble by Combining the Sets of Classification Rules

Jaydeep B. Patil[1], Vaishali Nandedkar[2]

[1]Savitribai Phule Pune University, Pune, Maharashtra, India

[2]Professor, Savitribai Phule Pune University, Pune, Maharashtra, India

Abstract: ***The task of data fusion is to identify the true values of data items (e.g., the true date of birth for Shivaji Maharaj) among multiple observed values drawn from different sources (e.g., Web sites) of varying reliability. The task of extracting knowledge from sample data is divided into a number of subtasks in case of machine learning applications. At some point, there is necessity to fuse or to combine the knowledge. And this knowledge is now "contained" in a number of classifiers in order to apply it to new data. It is impossible to exchange the raw data because of a limited communication bandwidth. Also, a central unit would constitute a single point of failure. In other data mining applications, knowledge extraction is split into subtasks due to memory or runtime limitations. Again, locally extracted knowledge must be consolidated later and quite often, the communication overhead should be low. Extracting information from multiple, possibly conflicting, data sources, and reconciling the values so the true values can be stored in a central data repository, is a problem of vital importance to the database and knowledge management communities.***

Keywords: Probabilistic logic, Bayesian methods, Covariance matrix, Knowledge engineering, Training, Data mining, classifier fusion, probabilistic classifier, Knowledge fusion, generative classifier, Coordinate measuring machines, Bayesian techniques, data mining.

1. Introduction

Extracting information from multiple, possibly conflicting , data sources, and reconciling the values so the true values can be storedin a central data repository, is a problem of vital importance to the database and knowledge management communities [a1].

In different machine learning applications, the assignment of concentrating information (e.g. grouping guidelines) from specimen information is separated into various subtasks. Commonplace illustrations are brilliant sensor systems, robot groups, or programming specialists that learn provincially in their surroundings. Sooner or later, there is need to circuit or to join the information that is currently "contained" in various classifiers with a specific end goal to apply it to new information. An application in the field of circulated interruption discovery in machine systems is depicted in[1]. Accordingly, it is difficult to trade the crude information in view of a constrained correspondence data transmission. Likewise, a focal unit would constitute a solitary purpose of disappointment. In other information mining applications, learning extraction is part into subtasks because of memory or runtime limits. Once more, generally concentrated learning must be solidified later and all the time, the correspondence overhead ought to beneath.

2. Related Work

On the off chance that we discuss "learning combination" we must be a great deal more exact to what sort of information we allude. Combination can happen at different levels or classes:

1) Data (e.g. sensor estimations or perceptions) or data removed from information can be melded to arrive at more certain conclusions, case in point.
2) Models or parts of models prepared from example information or data can be melded if the models were developed in an appropriated manner.
3) The yields of models can be intertwined, for instance, to get more certain choices or as on account of worldly and spatial information mining to infer conclusions for specific focuses in space and time.

Interestingly, the term "Bayesian knowledge fusion" is frequently connected with class one (see, e.g.,[7]). A few variations can be found in the writing, while the most fascinating ones are successive Bayesian estimation techniques[8] or the combination of a few probability works as on account of the autonomous probability pool approach. More subtle elements on this system which is truly different from our own as it addresses the first request and not the second request appropriations can be discovered in[9] and [10]; applications in the fields of interactive media , mechanical autonomy, or target recognition are illustrated in[11], [12], and [13].

Work in classification 3 breakers, for instance, the yield of " low-level " classifiers by averaging their labels[14] or utilizing their marks as information of a choice unit that could likewise be prepared from information (see, e.g.,[15],[16],[17]). More perplexing methodologies are packing or boosting (see, e.g.,[2]) which are frequently propelled by the thought that an outfit of "frail classifiers" may out perform a solitary classifier.

Work in classification 2 basically relies on upon the sort of information representation. Two primary fields can be distinguished: from one perspective, learning is regularly likened with requirements and there is some work concentrating on combination of demands, for example, [18], [19], [20]. Then again, information is regularly spoken to by graphical models that are liable to combination, for instance, Bayesian systems, (astute) point maps, or the like [21], [22], [23], [24], [25]. Few work deals with the mix of "building squares" of models [26]. There is one article that is almost related to our approach:[27] moreover depicts a Bayesian blend philosophy concentrated around hyper parameters and it in like manner ill-uses the thought of conjugate priors (cf. Section3.2) .This work, then again, is substantially more solid than the hyper parameter accord system concerning the determination of combination recipes & the application to classifier combination.

Likewise identified with class 2 are numerous endeavors to parallelize calculations, for example, desire boost methods for the parameterization of mixture models. The broadly utilized k- means bunching calculation is such a desire augmentation approach for a particular variation of Gaussian mixtures, for instance, all covariance grids are equivalent and products of the unit grid, mixture coefficients are dismissed, the task of information to model parts (spoke to by centroids) is double, not steady. Cases for parallelization methodologies can be found in [28] and [29] for example. It could even be demonstrated that correct methodologies are plausible as in they give the same comes about as though the information were not transformed in some imparted assets (e.g. imparted memory on account of multi-center architectures or dependable and for all time accessible correspondence foundation) and take into consideration a trade of transitional results with the comparing correspondence overhead. This work is expected to be utilized as a part of disseminated situations where correspondence expenses ought to be low.

Distributed intrusion detection: An application in the field of distributed intrusion detection in computer networks is described in [1]. Therefore, it is impossible to exchange the raw data because of a limited communication bandwidth. Also, a central unit would constitute a single point of failure. In other data mining applications, knowledge extraction is split into subtasks due to memory or runtime limitations.

Dirichlet distribution: According to [2], the conjugate prior of a multinomial is a Dirichlet distribution and the conjugate prior of a multivariate normal is a normal-Wishart distribution (also referred to as Gauss-Wishart).

Bayesian knowledge fusion: Interestingly, the term “Bayesian knowledge fusion” (which we also claim for our work) is often associated with category one (see, e.g., [7]).

Bayesian estimation Techniques: Several variants can be found in the literature, while the most interesting ones are sequential Bayesian estimation techniques [8] or the fusion of several likelihood functions as in the case of the independent likelihood pool approach.

Multimedia, robotics, or target detection are outline : More details on this technique— which is quite distinct from ours as it addresses the first order and not the second-order distributions—can be found in [9] and [10]; applications in the fields of multimedia, robotics, or target detection are outlined in [11], [12], and [13].

Work in category 3 fuses: The output of “low-level” classifiers by averaging their labels [14] or using their labels as input of a decision unit that could also be trained from data (see, e.g., [15], [16], [17]). More complex approaches are bagging or boosting (see, e.g., [2]) which are often motivated by the idea that an ensemble of “weak classifiers” may outperform a single classifier.

Work in category 2 essentially depends on the kind of knowledge representation. Two main fields can be identified: on the one hand, knowledge is often equated with constraints and there is some work focusing on fusion of constraints such as [18], [19], [20]. On the other hand, knowledge is often represented by graphical models that are subject to fusion, for example, Bayesian networks, (intelligent) topic maps, or the like [21], [22], [23], [24], [25].

Neural Networks: Number of work deals with this fusion connected with “building blocks” connected with designs, for instance, basis features when it comes to radial basis functionality nerve organs communities [26].

Bayesian fusion approach based on hyper parameters : There is one article that is closely related to our approach: [27] also describes a Bayesian fusion approach based on hyper parameters and it also exploits the concept of conjugate priors (cf. Section 3.2). This work, however, is much more concrete than the hyper parameter consensus technique concerning the derivation of fusion formulas and the application to classifier fusion.

Parallelization Approaches: Also related to category 2 are many attempts to parallelize algorithms such as expectation maximization techniques for the parameterization of mixture models. The widely used k-means clustering algorithm is such an expectation maximization approach for a very specific variant of Gaussian mixtures, for example, all covariance matrices are equal and multiples of the unit matrix, mixture coefficients are neglected, the assignment of data to model components (represented by centroids) is binary, not gradual. Examples for parallelization approaches can be found in [28] and [29] for instance.

Multicore Architectures: It could even be shown that exact approaches are feasible in the sense that they give the same results as if the data were not processed in distributed chunks [30]. These techniques typically assume some shared resources (e.g., shared memory in the case of multicore architectures or reliable and permanently available communication infrastructure) and allow for an exchange of intermediate results with the corresponding communication overhead. This work is intended to be used in distributed environments where communication costs should be low.

3. Problem Statement

In various machine learning applications, the task of extracting knowledge (e.g., classification rules) from sample data is divided into a number of subtasks. At some point, there is necessity to fuse or to combine the knowledge that is now “contained” in a number of classifiers in order to apply it to new data. It is impossible to exchange the raw data because of a limited communication bandwidth. Also, a central unit would constitute a single point of failure. In other data mining applications, knowledge extraction is split into subtasks due to memory or runtime limitations. Again, locally extracted knowledge must be consolidated later and quite often, the communication overhead should be low.

If we compare probabilistic generative classifiers to discriminative classifiers, discriminative classifiers are more likely to over fit to sample data as the (effective) number of parameters is typically quite high, or the classification performance is sometimes worse if data do not (at least nearly) meet the distribution assumptions. If two classifiers model similar processes they are likely to contain many similar components. So we introducing our new system for

- To detect such a situation to fuse all pairs of similar components.
- To generalize the approach to other distributions, in particular members of the exponential family of distributions and investigate how different prior distributions can be handled.
- And to find a more intuitive way to parameterize the fusion threshold is also a good challenge.

4. Methodological Foundations

Our System working on the basis of fundamental models these are.

1) **Input data:**
Classification rules are extracted from input sample data in a distributed way, it is necessary to combine or fuse these classification rules.

2) **Subtask:**
In various machine learning applications, the task of extracting knowledge (e.g., classification rules) from input sample data is divided into a number of subtasks. In data mining applications, knowledge extraction is split into subtasks due to memory or runtime limitations.

3) **Rule Fusion:**
At some point, there is necessity to fuse or to combine the knowledge that is now “contained” in a number of classifiers in order to apply it to new data.

Our fusion mechanism uses the hyper distributions obtained in the training process. Doing so, here retain these hyper distributions throughout the fusion process which has several advantages over a simple linear combination of classifiers parameters.

4) **Probabilistic Generative Classifier:**
The probabilistic classifiers offer the possibility to combine classifiers at the level of components of the mixture models (in the following these components are also referred to as “rules”). This can be accomplished by taking the union of all component sets and renormalizing the mixture coefficients.

Components or rules may be fused at the level of parameters. In this case, it is necessary to “average” the parameters of two or several components in an appropriate way if these components are regarded as being “sufficiently” similar.

a) **Classifier Ensemble**
First, the classifiers can be used in the form of ensembles, an idea for which a number of realizations exist (e.g., by weighting the classifiers depending on their classification performance). For probabilistic classifiers, where the outputs can be interpreted as posterior probabilities, this is quite simple.

b) **Combining Classification Rule**
If for a component of the first classifier no corresponding component of the second classifier exists (or vice versa), these components are simply combined in the resulting classifier. That is, the union of these component sets is built and the mixture coefficients are adapted accordingly.

5. Second Order Distribution

The key contribution of this work is that it shows an actual fusion of classifiers (or, components of classifiers) can be accomplished essentially by multiplying the second-order distributions if the classifier is based on certain members of the exponential family of distributions. Distributions such as Dirichlet or normal-Wishart distributions over parameters of the classifier.

6. Classification Rule

Output of the system is Knowledge represented by components of classifiers fused at a parameter level.

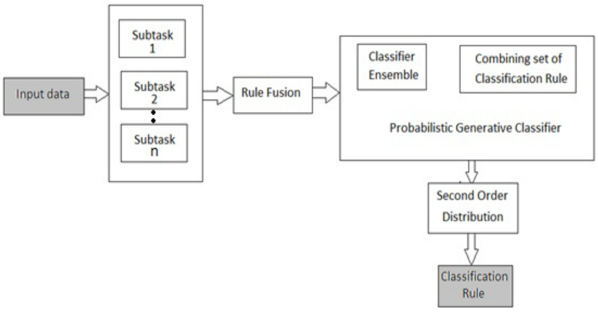
Input data
Subtask 1
Subtask 2
Subtask n
Rule Fusion
Classifier Ensemble
Combining set of Classification Rule
Probabilistic Generative Classifier
Second Order Distribution
Classification Rule

Figure 1: System Block Diagram

a) Probable methods of data analysis

The number of components and the classification performance of the overall classifiers obtained with the fusion/combination algorithm depend on the similarity threshold which has to be adjusted by the user depending on the application. Here analysis is with the developed fusion techniques for classifiers to artificial and real-world data sets.

Advantages of our system are

- The class posterior probabilities p(C/x) are very useful to weight single decisions when several classifiers are combined.
- A rejection criterion could easily be defined which allows to refuse a decision if none of the class posteriors reaches a pre-specified threshold or
- In dynamic environments it is possible to detect novel situations, for example, data that originate from new processes that did not exist when the initial training data were collected.

Possible drawbacks are

- These classifiers are more likely to over fit to sample data as the (effective) number of parameters is typically quite high, or
- The classification performance is sometimes worse if data do not (at least nearly) meet the distribution assumptions.

7. Conclusion

When two probablistic generative classifiers (CMM) fused into one, aCMMcomprises of a few segments each of which may thusly comprise of one multivariate typical dissemination demonstrating nonstop measurements of the inputs pace and numerous multinomial circulations.To distinguish parts of two classifiers that might be intertwined, we proposed asimilarity measure that works on the conveyances of the classifier. The expansion of insight combination methodology to more than two CMM classifiers is straight forward as it is conceivable to apply the strategy iteratively. It will surely be conceivable to utilize the same parameter values for all single combinations. In the event that the quantity of classifiers is known ahead of time it would likewise be conceivable to change the combination equations as needs be. The proposed methods could be utilized as a part of the field of circulated information mining, where information sets must be part to adapt to gigantic measures of information and where the correspondence expenses need to beneath. It is likewise conceivable to utilize the min conveyed situations where information are mainly handled as they emerge by regional standards .

Reference

[1] D. Fisch, M. Ja¨nicke, E. Kalkowski, and B. Sick, “Learning from Others: Exchange of Classification Rules in Intelligent Distributed Systems,” Artificial Intelligence, vol. 187-188, pp. 90-114, 2012.

[2] C.M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.

[3] D. Fisch, B. Ku¨ hbeck, B. Sick, and S.J. Ovaska, “So Near and Yet So Far: New Insight into Properties of Some Well-Known Classifier Paradigms,” Information Sciences, vol. 180, no. 18, pp. 3381-3401, 2010.

[4] N. Bouguila, “Hybrid Generative/Discriminative Approaches for Proportional Data Modeling and Classification,” IEEE Trans. Knowledge and Data Eng., vol. 24, no. 12, pp. 2184-2202, Dec. 2012.

[5] T.M. Hospedales, S. Gong, and T. Xiang, “Finding Rare Classes: Active Learning with Generative and Discriminative Models,” IEEE Trans. Knowledge and Data Eng., vol. 25, no. 2, pp. 374-386, Feb. 2013.

[6] D. Fisch, T. Gruber, and B. Sick, “SwiftRule: Mining Comprehensible Classification Rules for Time Series Analysis,” IEEE Trans. Knowledge and Data Eng., vol. 23, no. 5, pp. 774-787, May 2011.

[7] J. Sander and J. Beyerer, “Fusion Agents—Realizing Bayesian Fusion via a Local Approach,” Proc. IEEE Int’l Conf. Multisensor Fusion and Integration for Intelligent Systems, pp. 249-254, 2006.

[8] A. Makarenko and H. Durrant-Whyte, “Decentralized Data Fusion and Control in Active Sensor Networks,” Proc. Seventh Int. Conf. Information Fusion, pp. 479-486, 2004.

[9] O. Punska, “Bayesian Approaches to Multi-Sensor Data Fusion,”master’s thesis, Dept. of Eng., Univ. of Cambridge, 1999.

[10] H. Durrant-Whyte and T. Henderson, “Multisensor Data Fusion,”Springer Handbook of Robotics, B. Siciliano and O. Khatib, eds.chapter 25, pp. 585-610, Springer, 2008.

[11] L. Iocchi, N. Monekosso, D. Nardi, M. Nicolescu, P. Remagnino,and M. Valera, “Smart Monitoring of Complex Public Scenes,”Proc. Assoc. for the Advancement of Artificial Intelligence (AAAI) Fall Symp., 2011.

[12] P.K. Atrey and M.S. Kankanhalli, “Probability Fusion for Correlated Multimedia Streams,” Proc. ACM Int’l Conf. Multimedia,pp. 408-411, 2004.

[13] A. Barreiro, S. Liu, N. Namachchivaya, P. Sauer, and R. Sowers,“Data Assimilation in the Detection of Vortices,” Applications ofNonlinear Dynamics, Series Understanding Complex Systems, V.In, P. Longhini, and A. Palacios, eds., pp. 47-59, Springer, 2009.

[14] L.I. Kuncheva, “A Theoretical Study on Six Classifier FusionStrategies,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 2, pp. 281-286, Feb. 2002.

[15] M. Zhang, H. Song, S. Lv, Y. Li, X. Yu, and J. Bao, “Research on theMulti-Sensors Information Fusion Technique Based on the NeuralNetworks and Its Application,” Proc. Int’l Workshop Knowledge Discovery and Data Mining, pp. 93-96, 2009.

[16] B. Verma and A. Rahman, “Cluster Oriented Ensemble Classifier: Impact of Multi-Cluster Characterisation on Ensemble ClassifierLearning,” IEEE Trans. Knowledge and Data Eng., vol. 24, no. 4,pp. 605-618, Apr. 2011.

[17] X. Ceamanos, B. Waske, J.A. Benediktsson, J. Chanussot, M.Fauvel, and J.R. Sveinsson, “A Classifier Ensemble Based onFusion of Support Vector Machines for Classifying HyperspectralData,” Int’l J. Image and Data Fusion, vol. 1, no. 4, pp. 293-307, 2010.

[18] P. Gray et al., “KRAFT: Knowledge Fusion from Distributed Databases and Knowledge Bases,” Proc. Eighth Int. Workshop Database and Expert Systems Applications, pp. 682-691, 1997.

[19] K. ying Hui and P. Gray, “Constraint and Data Fusion in a Distributed Information System,” Proc. 16th British Nat’l Conf.Databases: Advances in Databases, pp. 181-182, 1998.

[20] K. ying Hui, “Knowledge Fusion and Constraint Solving in a Distributed Environment,” PhD dissertation, Dept. of ComputingScience, Univ. of Aberdeen, 2000.

[21] “A Multi Agent Systems Approach to Distributed Bayesian InformationFusion,” Information Fusion, vol. 11, no. 3, pp. 267-282, 2010.

[22] E. Santos Jr., J. Wilkinson, and E. Santos, “Bayesian Knowledge Fusion,” Proc. 22nd Int’l FLAIRS Conf., pp. 559-564, 2009.

[23] Y. Wang, B. Wu, and J. Hu, “A Semantic Knowledge Fusion Method Based on Topic Maps,” Proc. Workshop Intelligent Information Technology Application, pp. 74-76, 2007.

[24] H. Lu and B. Feng, “An Intelligent Topic Map-Based Approach to Detecting and Resolving Conflicts for Multi-Resource Knowledge Fusion,” Information Technology J., vol. 8, no. 8, pp. 1242-1248, 2009.

[25] A. Smirnov, M. Pashkin, N. Chilov, and T. Levashova, “KSNETApproach to Knowledge Fusion from Distributed Sources,”Computing and Informatics, vol. 22, no. 2, pp. 105-142, 2003.

[26] O. Buchtala and B. Sick, “Techniques for the Fusion of Symbolic Rules in Distributed Organic Systems,” Proc. IEEE Mountain Workshop Adaptive and Learning Systems, pp. 85-90, 2006.

[27] C.S.R. Fraser, L.F. Bertuccelli, H.-L. Choi, and J.P. How, “A Hyperparameter-Based Approach for Consensus under Uncertainties,”Proc. Am. Control Conf., pp. 3192-3197, 2010.

[28] A.G. Foina, J. Planas, R.M. Badia, and F.J. Ramirez-Fernandez, “Pmeans, A Parallel Clustering Algorithm for a HeterogeneousMulti-Processor Environment,” Proc. Int’l Conf. High Performance Computing and Simulation, pp. 239-248, 2011.

[29] Y. Li, K. Zhao, X. Chu, and J. Liu, “Speeding Up k-Means Algorithm by GPUs,” Proc. 10th IEEE Int’l Conf. Computer andInformation Technology, pp. 115-122, 2010.

[30] C.-T. Chu, S.K. Kim, Y.-A. Lin, Y. Yu, G. Bradski, A.Y. Ng, and K. Olukotun, “Map-Reduce for Machine Learning on Multicore,”Proc. Advances in Neural Information Processing Systems, pp. 281-288, 2006.

[31] D. Fisch, S.J. Ovaska, E. Kalkowski, and B. Sick, “In Your Interest - Objective Interestingness Measures for a Generative Classifier,”Proc. Third Int’l Conf. Agents and Artificial Intelligence, pp. 414-423, 2011.

[32] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification. John Wiley & Sons, 2001.

[33] D. Fisch, F. Kastl, and B. Sick, “Novelty-Aware Attack Recognition- Intrusion Detection with Organic Computing Techniques,” Proc.Seventh IFIP TC 10 Working Conf. Distributed, Parallel and Biologically Inspired Systems, pp. 242-253, 2010.