УДК 621.391

О.С. Высочина[1], С.И. Шматков[1], А.М. Салман[2]

[1]*Харьковский национальный университет имени В.Н. Каразина, Харьков*
[2]*Харьковский национальный университет радиоэлектроники, Харьков*

ОЦЕНКА ЭФФЕКТИВНОСТИ МЕТОДОВ КЛАССИФИКАЦИИ СОСТОЯНИЙ ТЕЛЕКОММУНИКАЦИОННОЙ СЕТИ

В статье оценивается эффективность различных алгоритмов классификации при распознавании состояния телекоммуникационной сети. Анализ алгоритмов классификации проводился при помощи системы анализа данных Weka. Исходным набором данных являлась динамика изменения оценок показателей качества телекоммуникационной сети. Анализ алгоритмов классификации показал, что наиболее эффективным методом классификации являются нейронные сети.

Ключевые слова: *телекоммуникационная сеть, методы классификации, нейронные сети.*

Введение

Постановка проблемы. Возрастающие темпы использования новых технологий информационного обеспечения приводят к увеличению количества сервисов, предоставляемых телекоммуникационной сетью. В связи с этим выдвигаются новые требования к качеству обслуживания. Для обеспечения требуемого качества обслуживания необходимо не только иметь информацию о текущем состоянии сети, но и уметь его прогнозировать. Таким образом, возникает необходимость в разработке новых методов управления телекоммуникационной сетью. Одной из компонент подобной системы управления является система мониторинга.

Системы мониторинга телекоммуникационных сетей присутствуют на рынке уже около восьми лет. Проведенный анализ показал, что подобные системы способны выдать пользователю статистику по ограниченному набору параметров сети, без учета их взаимосвязи. Поэтому для более полного анализа состояния сети в такие системы необходимо включать дополнительные модули обработки статистической информации, работа которых формально сводится к решению задачи классификации состояния телекоммуникационной сети.

Использование подобных модулей позволит оценить, а главное даст возможность спрогнозировать изменение показателей качества сети, с учетом взаимовлияния и доминирования информационных потоков.

Состояние телекоммуникационной сети характеризуется определенным набором показателей качества. Анализ динамики изменения основных показателей качества работы сети (среднепутевая задержка, джиттер, количество потерянных пакетов, среднее время простоя в очереди и т.д.) позволяет выделить наиболее типовые тренды этих величин. Конкретный вид зависимости определяется текущим состоянием сети, взаимным влиянием различных показателей качества между собой, внешними воздействиями и т.д. Таким образом, возможно по статистически полученной динамике изменения показателей качества распознавать и прогнозировать состояние сети.

Из вышесказанного можно сделать вывод, что задача выбора алгоритма классификации для распознавания состояния телекоммуникационной сети является **актуальной**.

Анализ последних исследований и публикаций. Для решения задачи классификации состояния телекоммуникационной сети существует множество разнообразных подходов и алгоритмов. Классические решения данной задачи рассматриваются в теории распознавания образов [1].

Работы [2 – 4] посвящены байесовской теории принятия решений, применением разделяющих функций и решением вопросов проверки гипотез. В работе [5] подробно рассмотрен метод потенциальных функций. В работе [6] особое внимание уделено статистической теории распознавания и методу "обобщенный портрет". Метод комитетов описан работе [7]. В работе [8] в качестве метода классификации предложен метод группового учета аргументов. Алгоритмы таксономии и анализа знаний представлены в работе [9]. В работе [10] предложены логические методы распознавания и поиска зависимостей. Многие работы в области теории распознавания и классификации связаны с применением искусственных нейронных сетей [11].

Проведенный анализ показал, что спектр методов достаточно широк, от классического статистического анализа до аппарата искусственных нейронных сетей. Решения, найденные различными алгоритмами, могут существенно отличаться друг от друга. Поиск наилучшего решения затруднен отсутствием общепризнанных универсальных критериев качества решений, поэтому для выбора наиболее эффективного метода использовалась система анализа данных Weka. Система анализа данных Weka представляет собой библио-

теку программ, реализующих линейные, комбинаторно-логические, статистические, нейросетевые, гибридные методы прогноза, классификации и извлечения знаний из прецедентов, а также коллективные методы прогноза и классификации [12].

Целью настоящей работы является оценка эффективности различных алгоритмов классификации состояний телекоммуникационной сети при помощи системы анализа данных Weka.

Изложение основного материала

Для сбора статистической информации о состоянии сети проведен натурный эксперимент на базе оборудования Cisco. Построен сегмент телекоммуникационной сети, состоящий из 5 маршрутизаторов и поддерживающий работу 100 абонентов. Сеть работала в штатном режиме (поддержка работы серверной базы данных, IP-телефонии, электронного документооборота). В процессе эксперимента при помощи SNMP-клиента с каждого маршрутизатора снимались значения показателей базы данных MIB [13], после чего выполнялось усреднение, и определялась корреляция между состоянием маршрутизатора и объемом трафика в сети.

На основе накопленной статистической информации выделено 6 типовых закономерностей изменения показателей качества телекоммуникационной сети. Динамика изменения основных показателей качества при различных состояниях сети носит вполне определенный характер. Основные зависимости интерполируются известными формульными соотношениями:

$$y = \frac{1}{(1+e^{-x})}, \quad D(f) = R, \quad E(f) = [0;1]; \qquad (1)$$

$$y = ax + b, \quad a, b \in R, \quad D(f) = R, \quad E(f) = R, \quad (a \neq 0), \quad E(f) = \{b\}\ (a = 0); \qquad (2)$$

$$y = e^{x}, \quad D(f) = R, \quad E(f) =]0; +\infty[; \qquad (3)$$

$$y = \ln x, \quad D(f) =]0; +\infty[, \quad E(f) = R; \qquad (4)$$

$$y = \log_a x, \quad a > 0, \quad a \neq 1, \quad D(f) =]0; +\infty[, \quad E(f) = R; \qquad (5)$$

где $D(f)$ – область определения функции $f : X \to R : X = D(f)$; $E(f)$ – множество значений функции $f : E(f) = \{f(x) \mid x \in X\} = f(X)$;

В процессе своего функционирования телекоммуникационная сеть, как сложная система, может прибывать в конечном множестве своих состояний S. В каждый конкретный момент времени она находится в конкретном k-м состоянии S_k, $(k = 1, 2, ..., z)$. Текущее состояние S_k телекоммуникационной сети определяется состоянием ее элементов $S_i^*, (i = 1, 2, ..., m)$. Количественно, состояния элементов характеризуются набором значений показателей качества x_j, $(j = 1, 2, ..., n)$. Формально состояния элементов могут быть представлены в виде множества векторов значений показателей качества:

$$\bar{S}_{i,l}^{*} = (x_1, x_2, ..., x_n), \qquad (6)$$

где $l = 1, 2, ..., t$ – множество возможных состояний элемента.

При рассмотрении каждого i-го элемента сети конкретное значение x_j является значением функции $y_j(S_i^*)$ его состояния.

При решении задачи классификации состояния сети в целом, значения показателей качества ее элементов считаются известными.

Таким образом, задача классификации состояния сети состоит в определении набора значений $y_j(S_i^*)$ для каждого элемента S_i^*. Следовательно, данная задача решается на основе задачи классификации состояния ее элементов и может быть представлена как специальная задача экстраполяции функции, зависящей от конечного числа разнотипных переменных – признаков, и заданной в виде таблицы значений в конечном числе точек.

Множество значений каждого признака (значения показателя качества) представляется функцией одной вещественной переменной, определенной на отрезке числовой оси и имеющей не более, чем заданное число точек разрыва первого рода.

Информация о текущем состоянии каждого элемента S_{il}^* задается набором из t строк: $y_1(S_{il}^*), y_2(S_{il}^*), ..., y_n(S_{il}^*)$ значений показателей качества a_{ji}. Такой вид представления информации о состоянии элементов сети позволяет сформировать матрицу обучения для каждого k-го состояния системы: $T_{nmt} = \left(y_j(S_{il}^*)\right)$.

$$\left.\begin{matrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ . & . & . & . \\ a_{m_1 1} & a_{m_1 2} & \cdots & a_{m_1 n} \end{matrix}\right\} \in S_1,$$

$$\left.\begin{matrix} a_{m_1+1,1} & a_{m_1+1,2} & \cdots & a_{m_1+1,n} \\ a_{m_1+2,1} & a_{m_1+2,2} & \cdots & a_{m_1+2,n} \\ . & . & . & . \\ a_{m_2 1} & a_{m_2 2} & \cdots & a_{m_2 n} \end{matrix}\right\} \in S_2, \qquad (7)$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\left.\begin{matrix} a_{m_{l-1}+1,1} & a_{m_{l-1}+1,2} & \cdots & a_{m_{l-1}+1,n} \\ a_{m_{l-1}+2,1} & a_{m_{l-1}+2,2} & \cdots & a_{m_{l-1}+2,n} \\ . & . & . & . \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{matrix}\right\} \in S_z.$$

Таким образом, для решения задачи классификации состояния элемента телекоммуникационной сети на основе таблицы обучения T_{nmt} требуется выбрать алгоритм, который на первом этапе по набору показателей качества будет корректно определять класс – состояние элемента сети; а на втором этапе – состояние всей сети, на основании набора состояний ее элементов.

В процессе работы телекоммуникационной сети показатели качества принимают значения, характеризующие текущее состояние сети. При изменении состояния сети наиболее чувствительные к причине изменения состояния показатели качества отходят от устоявшихся значений согласно своей динамике. Задача алгоритма классификации состоит в распознавании подобных ситуаций.

На вход алгоритма классификации поступают значения всех показателей качества, а результатом его работы является распознавание, какой из показателей качества динамически изменяется.

Рассмотрено более 60 алгоритмов классификации. Оценка эффективности работы алгоритмов производилась по вероятности правильного распознавания состояния телекоммуникационной сети.

Результаты проведенного вычислительного эксперимента и анализ литературы показали, что наибольшей эффективностью при решении задачи классификации состояния телекоммуникационной сети обладают следующие алгоритмы (рис. 1):

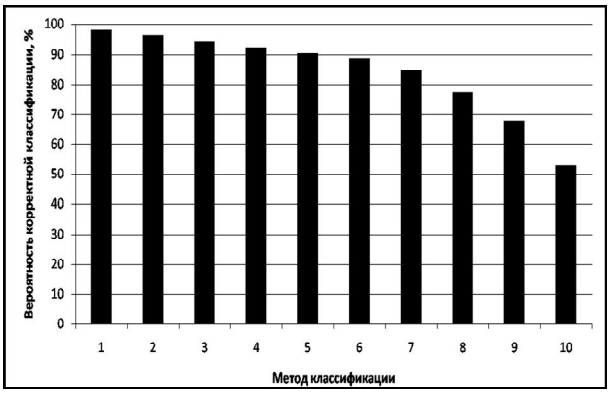

Рис. 1. Результаты работы наиболее эффективных алгоритмов классификации при распознавании состояния телекоммуникационной сети

- многослойный персептрон;
- нейросеть с радиальными базисными функциями;
- метод опорных векторов;
- метод "ближайшего соседа";
- наивно-байесовский подход;
- randomForest;
- алгоритм C4.8;
- метод байесовских сетей;
- алгоритм OneR;
- adaBoost.

Однако следует отметить у каждого из методов свои особенности применения и недостатки.

AdaBoost [14]

Склонность к переобучению при наличии значительного уровня шума в данных.

Требование достаточно длинных обучающих выборок.

Алгорим OneR [15]

Невозможность прямой обработки непрерывных переменных, данные необходимо преобразовывать в дискретный вид, что приводит к потере закономерностей.

Метод байесовских сетей [16]

Вычислительная сложность.

При попытке учесть большое количество зависимостей между переменными, оценки условных вероятностей приобретают большую дисперсию, таким образом, оценки параметров становятся недостоверными, что в итоге приводит к ухудшению качества классификации.

Ориентированность на обучающие данные из-за большого количества параметров, что приводит к хорошим результатам классификации на обучающих данных и неудовлетворительным результатам на тестовых данных, т.е. модель описывает не общие закономерности в структуре данных, а скорее набор частных случаев в обучающей выборке.

Алгоритм C4.8 [17]

Медленная работа на сверхбольших и зашумленных наборах данных.

RandomForest [18]

Построенная модель занимает большое количество памяти.

Склонность к переобучению.

Неспособность к экстраполяции.

Наивно-байесовский подход [19]

Невозможность непосредственной обработки непрерывных переменных – необходимо их преобразование к интервальной шкале, чтобы атрибуты были дискретными, однако подобные преобразования приводят к потере значимых закономерностей.

Влияние на результат классификации только индивидуальных значений входных переменных, комбинированное влияние пар или троек значений разных атрибутов не учитывается.

Метод "ближайшего соседа" [20]

Сложность выбора меры "близости" (метрики). От данной меры главным образом зависит объем множества записей, которые нужно хранить в памяти для достижения удовлетворительной классификации или прогноза.

Необходимость полного перебора обучающей выборки при распознавании, следствие этого – вычислительная трудоемкость.

Метод опорных векторов [21]

Для классификации используется не все множество образцов, а лишь их небольшая часть, которая находится на границах.

Выводы

При помощи системы анализа данных Weka произведен анализ алгоритмов классификации при распознавании состояния телекоммуникационной сети. Оценка эффективности работы производилась по вероятности правильного распознавания. Рассмотрено более 60 алгоритмов. Исходным набором данных являлась динамика изменения показателей качества телекоммуникационной сети. Анализ полученных результатов и литературы показали, что наилучшим образом при распознавании состояния телекоммуникационной сети себя проявил многослойный персептрон и нейронная сеть с радиальными базисными функциями [11], которые относятся к классу нейронных сетей. Поэтому можно сделать вывод о том, использование нейронных сетей при решении задачи классификации состояния телекоммуникационной сети является одним из наиболее эффективных методов. Поэтому перспективы дальнейших исследований в этом направлении связаны, в первую очередь, с выбором типа нейронной сети, и на основе этого с проведением более детальных модельных экспериментов.

Список литературы

1. Фукунага К. Введение в статистическую теорию распознавания образов / К. Фукунага. – М.: Наука, 1979. – 368 с.

2. Neyman J. On the Problem of the Most Efficient Tests of Statistical Hypotheses / J. Neyman, E.S Pearson. // Phil. Trans. R. Soc. – 1933. – Series A, №231. – P. 289-337.

3. Fisher R.A. The use of multiple measurements in taxonomic problems / R.A. Fisher // Ann. Eugenics. – 1936. – Part II, № 7. – P. 179-188.

4. Wald A. Contributions to the theory of statistical estimation and testing of hypotheses / A. Wald // Ann.Math.Stat. – 1939. – № 10. – P. 299-326.

5. Айзерман М.А. Метод потенциальных функций в теории обучения машин / М.А. Айзерман, Э.М. Браверманн, Л.И. Розоноэр. – М.: Наука, 1970. – 384 с.

6. Вапник В.Н. Теория распознавания образов / В.Н. Вапник, А.Я. Червоненкис. – М.: Наука, 1974. – 415 с.

7. Мазуров В.Д. Комитеты систем линейных неравенств / В.Д. Мазуров, М.Ю. Хачай // Автоматика и телемеханика. – 2004. – № 2. – С. 43-54.

8. Ивахненко А.Г. Системы эвристической самоорганизации в технической кибернетике / А.Г. Ивахненко. – К.: Техніка. – 1971. – 372 с.

9. Загоруйко Н.Г. Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. – Новосибирск: Изд–во Института математики, 1999. – 421 с

10. Лбов Г.С. Методы обработки разнотипных экспериментальных данных / Г.С. Лбов. – Новосибирск: Наука, 1981. – 160 с.

11. Хайкин С. Нейронные сети / С. Хайкин. – М.: Вильямс, 2006. – 1104 с.

12. Официальный сайт системы анализа данных Weka [Электронный ресурс]. – Режим доступа к сайту: http://www.cs.waikato.ac.nz/~ml/weka.

13. Семенов Ю.А. Протоколы Internet. Энциклопедия / Ю.А. Семенов. – М.: Горячая линия – Телеком, 2001. – 110 с.

14. Schapire Robert E. The boosting approach to machine learning: An overview / Robert E. Schapire // In MSRI Workshop on Nonlinear Estimation and Classification. – 2002. – 341 p.

15. Мурыгин К.В. Обнаружение объектов на изображении на основе каскада классификаторов / К.В. Мурыгин // Искусственный интеллект. – 2007. – № 2. – С. 104-108.

16. Тулупьев А.Л. Байесовские сети: логико-вероятностный подход / А.Л. Тулупьев, С.И. Николенко, А.В. Сироткин. – СПб.: Наука, 2006. – 607 с.

17. Lopez D. Error–correcting tree language inference / D. Lopez, S. Espana // Pattern Recognition Letters – 2002. – № 23. – P. 1-12.

18. Hastie. T. Random Forests / T Hastie, R. Tibshirani, J. Friedman // The Elements of Statistical Learning: Data Mining, Inference, and Prediction. – 2nd ed. – Springer–Verlag, 2009. – 746 p.

19. Brand E. Naive–Bayes and Nearest Neighbor / E. Brand, R. Gerritse. – BMS, 1998. – 513 p.

20. Дуда Р. Распознавание образов и анализ сцен / Р. Дуда, П. Харт. – М.: Мир, 1976. – 511 с.

21. Christopher J.C. Tutorial on Support Vector Machines for Pattern Recognition / J.C. Christopher, A. Burges // Data Mining and Knowledge Discovery 2, 1998 – P. 121-167.

Поступила в редколлегию 24.05.2010

Рецензент: д-р техн. наук, доц. А.В. Лемешко, Харьковский национальный университет радиоэлектроники, Харьков.

ОЦІНКА ЕФЕКТИВНОСТІ МЕТОДІВ КЛАСИФІКАЦІЇ СТАНІВ ТЕЛЕКОМУНІКАЦІЙНОЇ МЕРЕЖІ

О.С. Височина, С.І. Шматков, А.М. Салман

В статті оцінюється ефективність різних алгоритмів класифікації при розпізнаванні стану телекомунікаційної мережі. Аналіз алгоритмів класифікації проводився за допомогою системи аналізу даних Weka. Вхідним набором даних була динаміка зміни оцінок показників якості телекомунікаційної мережі. Аналіз алгоритмів класифікації показав, що найбільш ефективним методом класифікації є нейронні мережі.

Ключові слова: *телекомунікаційна мережа, методи класифікації, нейронні мережі.*

CLASSIFICATION METHODS EFFICIENCY ESTIMATION OF THE TELECOMMUNICATIONS NETWORK

O.S. Vysochyna, S.I. Shmatkov, A.M. Salman

In the present paper efficiency of different classification algorithms is estimated at recognition of the telecommunications network state. The analysis of classification algorithms was carried out by means of the Weka data analysis system. The initial data set was a dynamics of quality metrics estimations change of the telecommunications network. The analysis of classification algorithms showed that the most effective method of classification is neural network.

Keywords: *telecommunication network, methods of classification, networks of neurons.*